Exon-specific biomarkers in cancer:

Experimental validation of exon microarray data from colorectal and testicular cancers

Anne Cathrine Bakken

Department of Molecular Biosciences
Faculty of Mathematics and Natural Sciences
University of Oslo

Department of Cancer Prevention
Institute for Cancer Research
The Norwegian Radium Hospital, Rikshospitalet HF

Centre for Cancer Biomedicine
University of Oslo

A thesis for the Master's degree in Molecular Biosciences, December 2009

Acknowledgements

This work was carried out in the project Group of Genome Biology at the Department of Cancer Prevention, Rikshospitalet-Radiumhospitalet Medical Center, in the period March 2008 to December 2009.

First of all, I wish to thank my supervisor, Rolf I. Skotheim, for his great support throughout the project, and for always being positive and having good ideas. I would also like to thank my co-supervisor and head of the department, Ragnhild A. Lothe, for giving me the opportunity to being part of such an instructive group.

I am grateful to all of the group members, especially Guro for answering my TaqMan questions, Terje for helping me with the capillary electrophoresis, Anita and Sharm for providing me exon microarray data, and Deeqa and Marianne for our explanatory discussions and refreshing workouts in the backstairs.

Additionally, I wish to thank Morten Opsahl, my enthusiastic biology teacher at Fagerborg upper secondary school, who went beyond the curriculum and actually introduced me to the terms exons, introns and transcript diversity.

Finally, special thanks go to Benjamin, my family and friends, and the athletic milieu.

Oslo, December 2009

Anne Cathrine Bakken

Table of contents

ACKNOWLEDGEMENTS **2**

TABLE OF CONTENTS **3**

SUMMARY **5**

ABBREVIATIONS **7**

GENE SYMBOLS **9**

1. INTRODUCTION **10**

1.1 A BRIEF INTRODUCTION TO CANCER 10

1.2 TRANSCRIPT VARIATION IN CANCER 11

1.2.1 Alternative splicing *12*

1.2.2 Alternative use of promoters and polyadenlyation sites *17*

1.2.3 Global analysis of transcript variation *19*

1.3 TESTICULAR GERM CELL TUMOUR AS A MODEL FOR STEM-CELL RELATED MALIGNANCIES 20

1.4 COLORECTAL CANCER 23

2. AIMS **26**

3. MATERIALS AND METHODS **27**

3.1 MATERIALS 27

3.1.1 Cell lines *27*

3.1.2 Patient samples *27*

3.2 WHOLE-TRANSCRIPT EXPRESSION ANALYSIS 28

3.3 VALIDATION OF EXON MICROARRAY DATA 30

3.3.1 In silico exploration of candidate transcript variants *31*

3.3.2 Reverse transcriptase PCR (RT-PCR) *31*

3.3.3 Real-time RT-PCR 34

3.3.4 Detection of PCR products 36

3.3.5 DNA Sequencing 37

4. RESULTS 40

4.1 TRANSCRIPT VARIATION IN TESTICULAR GERM CELL TUMOUR 40

4.1.1 Polyamine-modulated factor 1 (PMF1) 41

4.1.2 DNA (cytosine-5)-methyltransferase 3β (DNMT3B) 44

4.1.3 Zinc finger protein 195 (ZNF195) 48

4.2 TRANSCRIPT VARIATION IN COLORECTAL CANCER 50

4.2.1 Splicing factor 1 (SF1) 51

4.2.2 DEAD (Asp-Glu-Ala-Asp) box polypeptide 17 (DDX17) 53

4.2.3 Solute carrier family 39 (zinc transporter), member 14 (SLC39A14) 55

5. DISCUSSION 59

5.1 TRANSCRIPT VARIATION IN TESTICULAR GERM CELL TUMOUR 60

5.2 TRANSCRIPT VARIATION IN COLORECTAL CARCINOMA 61

5.3 METHODOLOGICAL CONSIDERATIONS – ASSAY DESIGN 63

5.4 METHODOLOGICAL CONSIDERATIONS – DETECTION OF PCR PRODUCTS 68

5.5 VALIDATION FREQUENCY 69

6. CONCLUSIONS 72

7. FUTURE STUDIES 74

8. REFERENCE LIST 76

APPENDIX I – PRIMER AND PROBE INFORMATION 81

Summary

Aberrant splicing of pre-mRNA is common in cancer cells, and may create cancer-specific transcript variants. These abnormal transcripts, as well as their derived protein products, may have the potential to function as diagnostic and prognostic biomarkers, and even as therapeutic drug targets. Current biomarkers do not satisfy the requirements of sensitivity and specificity; accordingly, the search for novel transcript structures in cancer is promising.

In the present thesis, the overall aim was to discover novel cancer-specific transcript variants, as well as skewed ratios between pre-existing transcripts, in colorectal and testicular cancers. To reach that goal, candidate differentially expressed exons between malignant and non-malignant samples from exon-level microarray data were investigated. As the results from such microarrays are not intuitive in identifying the transcript structures, and numerous false positives arise, it is absolutely essential to validate and explore the exon microarray results by another method.

We used RT-PCR based approaches on the 30 apparently most promising candidate genes identified by exon microarrays to validate possible differential exon inclusion, as well as to reveal the structure of the candidate cancer-specific transcript variants. Six of the 30 candidate genes (*SLC39A14*, *DDX17*, *SF1*, *ZNF195*, *PMF1*, and *DNMT3B*) were successfully validated as having differences in transcript variants in the same samples as analysed by the exon microarray analyses. These transcripts were further investigated in extended sample series. Transcript variants of *SLC39A14*, *DDX17* and *SF1*, originally identified from the colorectal cancer dataset, were further investigated in 105 colorectal cancers, eight normal tissues, and in six colon cancer cell lines. Transcript variants of *ZNF195*, *PMF1* and *DNMT3B*, originally identified from the testicular cancer dataset, were further investigated in 25 testicular germ cell tumours, as well as in six pre-malignant intratubular germ cell neoplasias, and five normal testicular parenchyma tissue samples, in addition to five embryonal carcinoma cell lines and two embryonic stem cell lines.

Differential ratios between pre-existing transcripts were found for all six genes. *ZNF195* was identified as a particularly highly expressed gene in pluripotent samples (both embryonal carcinoma and embryonic stem cells), and with a malignancy-specific ratio between the transcript variants. Regarding *DDX17* and *SF1*, the colorectal cancer tissue samples revealed a higher proportion of the candidate cancer-specific transcript as compared to the normal tissue samples. However, the most promising candidate biomarker found in this study was *SLC39A14*, which appeared to have two mutually exclusive exons with differential inclusion in malignant and normal tissues from colon and rectum. Real-time RT-PCR revealed that 78 of the 105 (74 %) colorectal cancer tissue samples exclusively expressed the cancer-specific exon, whereas six of the eight normal tissue samples exclusively expressed the other transcript variant. Including the remaining samples, only two of the 105 colorectal cancer tissue samples showed an expression ratio between the two different transcripts that mixed with the normal tissue samples.

Altogether, we have identified novel transcript variants that may serve as improved biomarkers in colorectal (*SLC39A14*) and testicular cancer (*ZNF195*).

Abbreviations

3'SS	3' splice site
3'UTR	3' untranslated region
5'SS	5' splice site
5'UTR	5' untranslated region
bp	Base pairs
APS	Alternative polyadenylation site
BPS	Branch point sequence
cDNA	Complementary DNA
CIMP	CpG island methylator phenotype
CIN	Chromosome instable
CML	Chronic myelogenous leukaemia
CRC	Colorectal cancer
Ct	Cycle threshold
ddNTP	Dideoxyribonucleotide triphosphate
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
EC	Embryonal carcinoma
EDTA	Ethylenediaminetetraacetic acid
ES	Embryonic stem cell
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
EST	Expressed sequence tag
FAP	Familial adenomatous polyposis
FDR	False discovery rate
FIRMA	Finding isoforms using robust multichip analysis
Hi-Di formamide	Highly deionised formamide
HNPCC	Hereditary non-polyposis colorectal cancer
hnRNP	Heterogenous nuclear ribonucleoprotein
i(12p)	Isochromosome 12p
ICM	Inner cell mass
IGCN	Intratubular germ cell neoplasia
ISE	Intronic splicing enhancer
ISS	Intronic splicing silencer
MGB	Minor groove binder
miRNA	Micro ribonucleic acid
MMR	Mismatch repair
mRNA	Messenger RNA
MSI	Microsatellite instable
NMD	Nonsense-mediated mRNA decay
PCR	Polymerase chain reaction
PGC	Primordial germ cell
PPT	Polypyrimidine tract
pre-mRNA	Precursor messenger ribonucleic acid
PSR	Probe selection region
RACE	Rapid amplification of cDNA ends
RNA	Ribonucleic acid
RT-PCR	Reverse transcriptase polymerase chain reaction
siRNA	Short interfering RNA
snRNA	Small nuclear ribonucleic acid
snRNP	Small nuclear ribonucleic protein

SR protein	Serine/arginine-rich protein
TAE buffer	Tris-acetate EDTA buffer
TGCT	Testicular germ cell tumour
Tm	Melting temperature
TSS	Transcription start site
UHR	Universal human reference

Gene symbols

ABL	c-abl oncogene 1
APC	Adenomatous polyposis coli
BCL-X	BCL2-like 1
BCR1	Breakpoint cluster region
BRCA1	Breast cancer 1, early onset
DDX17	DEAD (Asp-Glu-Ala-Asp) box polypeptide 17
DNMT3B	DNA (cytosine-5)-methyltransferase 3β
GCNT3	Glucosaminyl (N-acetyl) transferase 3, mucin type
ING4	Inhibitor of growth family, member 4
LEF1	Lymphoid enhancer-binding factor 1
MLH1	mutL homolog 1, colon cancer, nonpolyposis type 2 (*E. coli*)
MSH2	mutS homolog 2, colon cancer, nonpolyposis type 1 (*E. coli*)
PMF1	Polyamine-modulated factor 1
SF1	Splicing factor 1
SLC39A14	Solute carrier family 39 (zinc transporter), member 14
ZNF195	Zinc finger protein 195

1. Introduction

1.1 A brief introduction to cancer

Malignant tumours are generally considered to arise from a single cell of origin [1], but there is evidence for the existence of polyclonal tumours as well [2]. Tumourigenesis is a multistep process resulting from acquired genetic and epigenetic variability among the cancer cells, leading to the sequential selection of more aggressive sublines [1,3]. The following phenotypic changes have been referred to as "the hallmarks of cancer": self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis [4]. Genomic instability [5] and the ability to escape the host immune system [6] are also regarded as acquired capabilities of cancer.

As early as in 1914, the German pathologist Theodor Boveri published a reflective monograph on the connection between chromosomal anomalies and tumour development [7]. Almost half a century passed before Boveri's idea received some validation. In 1960, two cytologists working in Philadelphia noted that an abnormal, unusually small chromosome was characteristically present in chronic myelogenous leukaemia (CML) cells [8]. The origin of this so-called Philadelphia chromosome was not solved until 1973, when Janet D. Rowley demonstrated that a reciprocal translocation between chromosomes 9 and 22 was responsible for creating the minute chromosome [9].

Today we know that this particular translocation is present in approximately 95 % of all CML cases [10], and that it results in the fusion of the *BCR1* gene and the proto-oncogene *ABL*. The hybrid BCR1-ABL oncoprotein product functions as a constitutively activated tyrosine kinase [11].

Proto-oncogenes encode proteins which functions often are to control cell proliferation, apoptosis, or both, and as intimated above, cellular proto-oncogenes can be activated by structural alterations resulting from gene fusions. Mutations affecting crucial residues, and overexpression caused by *e. g.* genomic amplifications [12] or juxtaposition to enhancer elements [13], represent other ways to activate oncogenes. Inactivations of tumour suppressor genes, in addition to alterations of DNA repair genes (the caretakers), also contribute to growth and proliferation advantages and stepwise tumour formation. Tumour suppressor genes encode proteins whose normal function is to reduce the likelihood of tumour formation, and usually both alleles need to be inactivated to provide tumourigenic effects. The loss of tumour suppressor gene function can occur either through genetic mutation or epigenetic silencing of genes via promoter methylation. Despite that the genome in most cancer cells is hypomethylated at repetitive DNA sequences at the global level, local hypermethylation of CpG islands[1] in the promoter regions of tumour suppressor genes is a major event in the origin of many cancers [3].

The first drafts of the human genome was published in 2001 [14,15], and the current estimate of human genes is ranging from 20 000 to 25 000 [16]. The variation at the RNA and protein levels accounts for the high diversity and complexity of higher eukaryotes. Instability at the transcriptome and proteome levels adds further complexity to cancer cells.

1.2 Transcript variation in cancer

One mechanism that accounts for the greater complexity and protein diversity of higher eukaryotes is alternative splicing of the RNA molecules. Through alternative splicing, a single gene is able to generate several transcript variants from one type of precursor messenger RNA (pre-mRNA), which may lead to the production of

[1] CpG island – region of the genome with significantly higher than average content of the CpG dinucleotide.

different protein isoforms. Alteration of the normal process of alternative splicing is frequent in cancer cells, and may result in the production of previously absent mRNAs or in the modification of the ratio between pre-existing mRNAs. Both these scenarios may be used as diagnostic and prognostic biomarkers, although the most promising biomarkers are those that are entirely cancer-specific. Furthermore, aberrant splice products with functional implications may be exploited in targeted therapeutics.

Most studies of cancer transcriptomes have been purely quantitative in nature, and do not take into account the qualitative variation of transcript variants. In the following, the contribution of alternative splicing and alternative use of promoters and polyadenylation sites to transcriptome diversity will be explained – both in general and in relation to cancer. Fusion transcripts[2] are also common in cancer [17], but will not be further described in this thesis. Finally, methods used in the analysis of transcriptomes will be discussed.

1.2.1 Alternative splicing

The median number of exons in a human protein-coding gene is seven, while the median internal exon length is 122 bp [14]. The exons are interrupted by usually much longer non-coding introns. Genes are transcribed into pre-mRNAs, containing exons as well as introns, and in the process of splicing the introns are removed from the transcripts and the adjacent exons are ligated.

Splicing of metazoan pre-mRNAs involves two transesterification reactions and functional groups from three poorly conserved reactive regions in the primary transcript; the 5’ and 3’ splice sites (5’SS and 3’SS, respectively), and the branch point sequence (BPS) (Figure 1) [as thoroughly reviewed in reference 18].

[2] Fusion transcript – a chimeric RNA encoded by a fusion gene or formed by trans-splicing of two different genes.

Figure 1. Pre-mRNA reactive regions. The locations of the three pre-mRNA reactive regions, the 5' and 3' splice sites (5'SS and 3'SS, respectively) and the branch point sequence (BPS), are indicated. The polypyrimidine tract (PPT) is located between the BPS and the 3'SS. The BPS, PPT and 3'SS sequence elements are all situated in the 3' end of the intron.

First, the phosphodiester bond at the 5'SS is attacked by the 2'-hydroxyl of an adenosine of the intronic BPS, which generates a free 5' exon and an intron lariat-3' exon. Then, the 3'-hydroxyl of the 5' exon attacks the phosphodiester bond at the 3'SS, resulting in exon ligation and excision of the lariat intron. The folding of pre-mRNA introns in a manner consistent with splicing is dependent on several *trans*-acting factors that comprise the spliceosome [18].

The splicing reaction is carried out in the nucleus by the spliceosomal complex consisting of five small nuclear ribonucleoproteins (snRNP), U1, U2, U4/U6, and U5, and more than 100 other proteins [19]. The spliceosome recognises the 5'SS and the 3'SS, in addition to the BPS and the polypyrimidine tract (PPT), and catalyses the cut-and-paste reactions. Assembly of the spliceosome on the pre-mRNA begins with the binding of U1 snRNP to the 5'SS. The earliest assembly phase also involves binding of the interacting proteins SF1/BBP and U2 auxiliary factor (U2AF) to the BPS, and the PPT and 3'SS sequences, respectively. Together, these molecular interactions yield the spliceosomal E complex. Subsequently, the U2 snRNP binds the BPS, leading to the displacement of SF1/BBP and the formation of the A complex. Then, the preassembled tri-snRNP consisting of U4/U6 and U5 snRNPs are recruited to the transcript, resulting in the formation of the B complex. During the activation of the spliceosome the complex goes through extensive structural and compositional rearrangements, including U1 and U4 destabilisation or release, resulting in the formation of the catalytically active spliceosome [18].

Alternative splicing can be defined as the joining of different 5' and 3' splice sites to produce distinct mRNAs [20], and has been estimated to occur in 86 % of all human genes [21]. Recognition and selection of splice sites in higher eukaryotes is

influenced by pre-mRNA regulatory sequences known as intronic and exonic splicing enhancers and silencers (abbreviated ISE, ESE, ISS, and ESS). These *cis*-acting elements have positive (ISE and ESE) or negative (ISS and ESS) influence on splice-site usage, and mediate their effects mainly by binding *trans*-acting regulatory factors [18] (Figure 2).

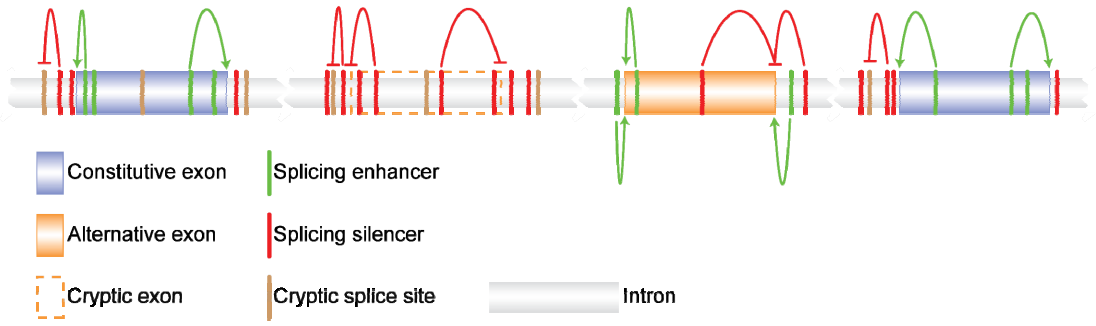


Figure 2. Cis-acting elements involved in constitutive and alternative splicing. Green arrows indicate the positive activity of splicing enhancers on the selection of adjacent splice sites in the alternative and constitutive exons. Red arrows with flat heads signify the negative activity of splicing silencers in cryptic exons and intronic regions proximal to cryptic splice sites, as well as in the alternative exon [modified from reference 22].

The *trans*-acting factors recruit the snRNP subunits of the splicing machinery to the adjacent splice site or, in the case of negative regulators, prevent their association. ESEs are often bound by proteins of the serine-arginine-rich (SR) protein family, whereas ESSs are typically bound by heterogeneous nuclear RNPs (hnRNPs). Eventually, it is the sum of multiple factors that decides whether a particular site is recognised by the spliceosome for inclusion of the adjacent exon in the mature mRNA [18].

The regulation of alternative splicing depends not only on the interaction between *cis*-acting elements and *trans*-acting factors, but also on the rate and pausing of transcriptional elongation. This can be explained by the fact that in the absence of internal stops, or during fast elongation, the 3'SSs of both an alternative cassette exon and a downstream constitutive exon might be presented simultaneously to the splicing machinery. Usually, the 3'SS of the alternative exon is weaker than the 3'SS of the constitutive exon, resulting in exon skipping. In the opposite, when the RNA polymerase II halts, or moves slowly, the weak 3'SS of the alternative exon is

available to the spliceosome on its own, leading to inclusion of the alternative exon [23].

The most common type of alternative splicing involves cassette-type alternative exons, which are either skipped or included in the mature mRNA. Alternative selection of 5' and 3' splice sites within exon sequences are also frequent, resulting in varying inclusion of sequence from a particular exon. Other types of alternative splicing events include intron retention and exons that are spliced in a mutually exclusive fashion. Mutually exclusive exons are never expressed in the same mature mRNA [22] (Figure 3).

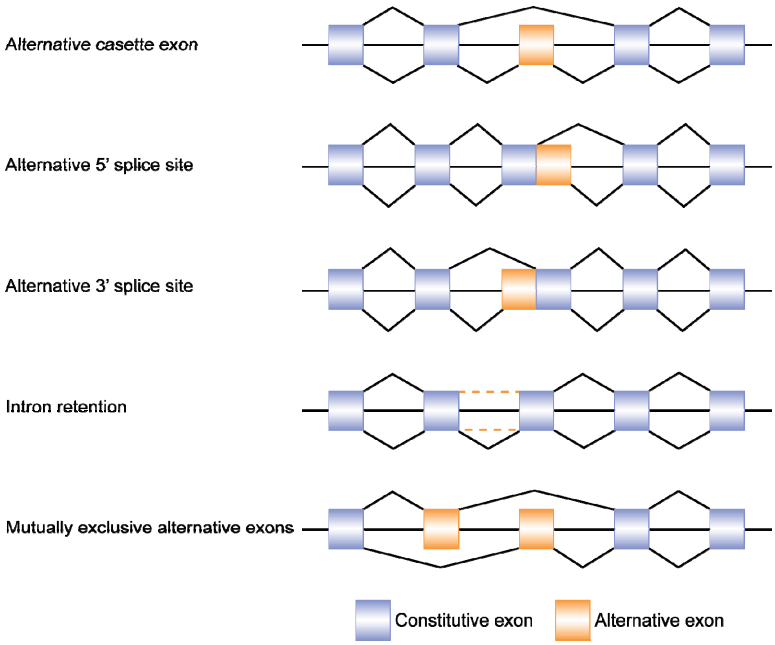


Figure 3. Different alternative splicing events. Blue and orange boxes indicate constitutive and alternative exons, respectively. These colour codes are recurrent in several figures in this thesis.

Up to one-third of human alternative splicing events create a premature stop codon, that generally results in the degradation of the transcript by nonsense-mediated mRNA decay[3] (NMD) [25]. However, some alternative mRNA species can be translated into different protein isoforms with potentially tumourigenic effects. Both

[3] Nonsense-mediated mRNA decay (NMD) – a pathway that degrades mRNAs that have premature termination codons [24].

mutations in *cis*-splicing regulatory elements and alterations of *trans*-splicing factors may result in abnormal splicing (Figure 4) [20]. Several exonic nonsense mutations[4] are linked to exon skipping [24], and this is the case for the breast-cancer susceptibility gene *BRCA1*(Figure 4A). An inherited mutation of an ESE element in *BRCA1* prevents binding of ASF/SF2, leading to inappropriate skipping of the cassette exon harbouring the mutant ESE [26]. Point mutations, including synonymous mutations[5], in coding as well as in non-coding regions, may affect splice sites and splicing regulatory sequences. Consequently, it is important to consider these mutations as well concerning alternative splicing.

Changes in the concentration, localisation, composition, or activity of *trans*-acting regulatory factors may also result in aberrant splicing. For example, the phosphorylation status of SR proteins affects the splicing of the *BCL-X* pre-mRNA. Alternative splicing of this gene results in two isoforms encoding proteins with antagonistic functions; a long antiapoptotic isoform (BCL-XL) and a short proapoptotic isoform (BCL-XS), with the long isoform being aberrantly overexpressed in several tumours (Figure 4B) [27].

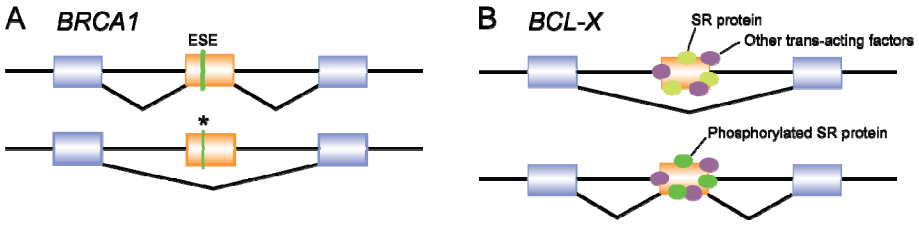


Figure 4. Mutations in cis-splicing regulatory elements and alterations of trans-splicing factors. (A) A mutation (*) in an exonic splicing enhancer (ESE) of *BRCA1* prevents the binding of ASF/SF2 to the immature transcript, leading to inappropriate exon skipping. **(B)** The phosphorylation status of SR proteins affects the activity of these proteins and thus the ratio between proapoptotic (exon skipping) and antiapoptotic (exon inclusion) BCL protein [modified from reference 27].

[4] Nonsense mutation – any change in DNA that replaces a codon specifying an amino acid with a translation-termination codon.

[5] Synonymous mutation – does not change the amino acid encoded by the triplet.

1.2.2 Alternative use of promoters and polyadenlyation sites

The use of alternative promoters and alternative polyadenylation sites also contribute to the complexity of the proteome. Most genes have multiple core promoters[6] within which there are multiple transcription start sites (TSSs)[7] [28]. The use of alternative promoters enables diversification of transcriptional regulation within a single locus, thus playing an important role in the control of gene expression [29]. In Figure 5 A and B, two more consequences of alternative promoter usage are depicted. The differential use of distinct promoters, with no translation initiation site between them, results in the transcription of mRNAs with heterogeneous 5' untranslated regions (5'UTR) coding identical proteins (Figure 5A). The distinct 5'UTRs, however, may affect the stability or translation efficiency of the mRNAs. If, on the other hand, there is a translation initiation site directly downstream of each of the possible promoters, mRNAs encoding distinct proteins can be generated (Figure 5B). Possible mechanisms of how a specific promoter is used in preference to others include core promoter structure, concentration of *cis*-regulatory elements, and regional epigenetic modifications [29].

Several oncogenes and tumour suppressor genes have multiple promoters, and aberrant promoter usage is in some cases directly linked to cancerous cell growth [30,31]. For example, *LEF1*, which encodes proteins mediating transcriptional regulation of WNT/β-catenin target genes, is transcribed by two alternative promoters. The upstream promoter produces full-length LEF1 protein, which recruits β-catenin to WNT target genes, whereas the other promoter derives a truncated form that cannot interact with β-catenin and instead suppresses WNT regulation of target genes. The promoter that produces full-length, growth-promoting LEF1 is aberrantly activated in colon cancer tissues [30].

[6] Core promoter – the genomic region that surrounds a TSS or cluster of TSSs, and that is required to recruit the transcription initiation complex and initiate the transcription.

[7] Transcription start site (TSS) – a nucleotide in the genome that is the first to be transcribed into a particular RNA.

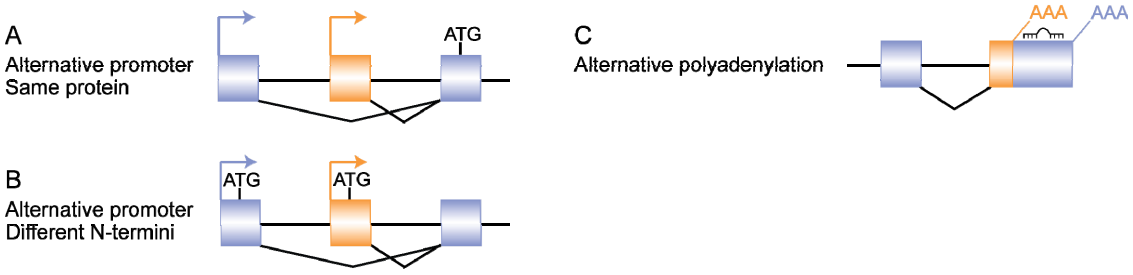


Figure 5. Alternative promoters and polyadenylation. (A) The use of the alternative promoter (orange arrow), rather than the upstream promoter, results in the transcription of a shorter pre-mRNA, and in the use of an alternative first exon (orange box). However, the translation initiation site is downstream of both promoters, and the resulting protein products will accordingly be identical. **(B)** The different pre-mRNAs are translated into proteins differing in their N-termini. **(C)** Recognition and utilization of a proximal polyadenylation site, leading to a shortening of the 3'UTR. Micro-RNA binding sites in metazoan mRNAs usually lie in the 3'UTR, as indicated by a partially hybridising oligonucleotide. The use of possible alternative polyadenylation sites upstream of the depicted 3'UTR, leading to a truncated protein product with a different C-terminus, is not shown.

Cleavage and polyadenylation of the pre-mRNA 3' ends are required for the maturation of most transcripts. The pre-mRNA is cleaved ten to 30 nucleotides after the polyadenylation signal and a poly(A) tail is then added. A strong polyadenylation site is usually located at the end of the 3'UTR. However, nearly all genes have additional polyadenylation signals in their 3'UTR (Figure 5C), and differential expression of *trans*-acting factors may explain the use of proximal polyadenylation sites. Shortening of the 3'UTR by alternative cleavage and polyadenylation may influence mRNA nuclear export and cytoplasmic localisation, as well as mRNA stability and translational efficiency [32].

Micro-RNAs[8] (miRNAs) interact with the 3'UTR of mRNAs and affect protein synthesis by mRNA destabilisation and translational repression [33]. Recently, Mayr and Bartel compared similarly proliferating nontransformed cell lines and cancer cell lines, and found that the cancer cell lines often expressed substantial amounts of mRNA isoforms with shorter 3'UTRs [32]. These shorter isoforms usually resulted from alternative cleavage and polyadenylation, and exhibited increased stability and typically produced a ten-fold more protein, partially due to the loss of miRNA-mediated repression. In particular, they observed that the expression of the shorter

[8] Micro RNA – endogenous, non-coding RNA, approximately 23 nucleotides, important in post-transcriptional gene regulation.

mRNA isoform of the proto-oncogene *IMP1*, an RNA binding protein that is overexpressed in a variety of human cancers, led to far more oncogenic transformation than did expression of the full-length annotated mRNA. Alternative cleavage and polyadenylation should therefore be considered mechanisms to activate cellular proto-oncogenes.

1.2.3 Global analysis of transcript variation

Bioinformatic analyses of expressed sequence tag (EST)[9] data have provided a rich source of information for the identification and analysis of transcript diversity. However, EST coverage is typically biased toward the 3' and 5' ends of transcripts, and in general there are insufficient numbers of sequenced transcripts to infer the frequency of specific events, like inclusion or skipping of an alternative cassette exon in a given tissue.

Some of the limitations inherent in the analysis of ESTs have been overcome by the development of DNA microarrays. Historically, microarrays have interrogated the 3' ends of RNAs, and used expression at the 3' end to approximate expression of the entire gene. However, this approach assumes that the 3' end of each gene is clearly defined, and does not discriminate between distinct transcripts, as opposed to whole-transcript expression analysis with exon microarrays or exon-exon junction microarrays.

Prior to the work of the present thesis, the transcriptomes of colorectal carcinomas and testicular germ cell tumours, with their normal counterparts, were analysed by the GeneChip Exon 1.0 ST Array (Affymetrix, Santa Clara, CA, USA) [35]. This microarray contains approximately 5.4 million probes grouped into 1.4 million probe sets, which are scattered throughout the genome examining more than one million known and predicted exons. A probe selection region (PSR) is the target sequence for

[9] Expressed sequence tag (EST) – two to eight hundred nucleotide sequence generated from either the 5' or 3' end of a cDNA clone [34].

which a probe set is designed and represents a region of the genome that is predicted to act as a single unit of transcriptional behaviour[10].

The ability to measure exon-level differences in gene expression enables indications of differential splicing, promoter usage and polyadenylation, as well as existence of fusion transcripts. Additionally, results from exon microarray data may indicate the existence of novel transcripts, which is an advantage over exon-exon junction microarrays [36], that are typically designed against observed or annotated junctions. Deep sequencing, also known as high-throughput sequencing or next-generation sequencing, of complementary DNA (cDNA) fragments has also been performed in the analysis of transcriptomes and is a powerful method for identification of novel transcripts [21,37].

1.3 Testicular germ cell tumour as a model for stem-cell related malignancies

Testicular cancer is the most common form of cancer among young adult men in the Western world[11]. These patients have a relatively favourable prognosis, with a five-year survival rate of about 95 %, due to the extreme chemosensitivity of testicular cancers.

The great majority (more than 95 %) of testicular tumours are of germ cell origin, and are therefore called testicular germ cell tumours (TGCT). Three epidemiologically, clinically and histologically distinct forms of TGCTs have been described; the teratomas and yolk-sac tumours of neonates and infants (type 1), the spermatocytic seminomas of elderly men (type 3), and the type 2 seminomas and non-seminomas of adolescents and young adults, which is the most common [38].

[10] Often, each PSR is an exon, however, due to potentially overlapping exon structures or alternative splice site usage, several PSRs may form contiguous, non-overlapping subsets of a true biological exon.

[11] http://seer.cancer.gov

The type 2 TGCTs develop after puberty from pre-malignant, non-invasive intratubular germ cell neoplasia (IGCN) [39], which is thought to arise during foetal development from a primordial germ cell (PGC) or gonocyte [40]. IGCN has the ability to develop into two histologically distinct tumours, seminomas and non-seminomas. The seminomas are composed of a uniform population of undifferentiated cells, which resemble the PGC, whereas the heterogeneous group of non-seminomas consists of embryonal carcinoma (EC), choriocarcinomas, yolk sac tumours, and teratomas (Figure 6). The pluripotent EC cells can be regarded as the stem cell population of the non-seminomas, as they have the ability to differentiate into the other histological subgroups of non-seminomas [41]. Understanding the biology of EC cells may thus bring knowledge not only about TGCTs, but also concerning cancer stem cells[12].

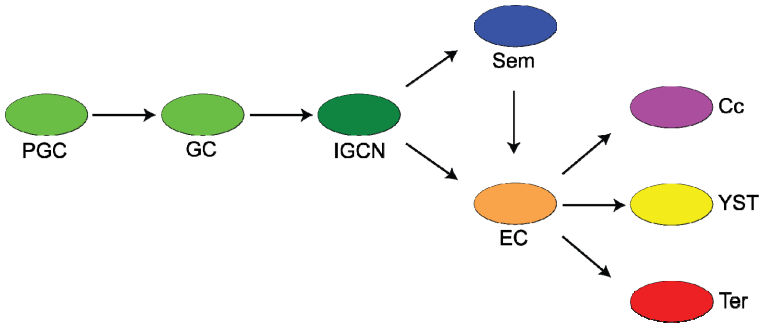


Figure 6. Development of TGCT. It is thought that intratubular germ cell neoplasia (IGCN) arises during foetal development by malignant transformation of a primordial germ cell (PGC) or gonocyte (GC). After puberty, IGCN may develop into invasive cancer in the form of seminoma (Sem) or embryonal carcinoma (EC) histological subtypes. The pluripotent EC cells have the potential to differentiate into distinct extra-embryonic tissues, like choriocarcinomas (Cc) and yolk sac tumours (YST), in addition to embryonic and somatic tissues (teratomas, Ter). Together, EC, Cc, YST, and Ter constitute the heterogeneous group of non-seminomas. The origin of non-seminomas is somewhat uncertain; either, embryonal carcinomas develop directly from IGCN, or they develop through a seminoma stage [43].

More than 80 % of TGCT genomes are characterised by the presence of isochromosome 12p[13] [44], i(12p), and most i(12p) negative TGCTs have

[12] Cancer stem cell – a cell within a tumour that possesses the capacity to self-renew and to cause the heterogeneous lineages of cancer cells that comprise the tumour [42].

[13] Isochromosome – an abnormal chromosome with two identical arms due to duplication of one arm and loss of the other.

amplifications of 12p sequence. Additionally, the general TGCT genome is hypo- to hypertriploid with extensive chromosome instability, and has erased parental imprinting. Furthermore, seminomas and non-seminomas have distinct epigenetic phenotypes. Whereas CpG island methylation is virtually absent in seminomas, the degree of methylation in non-seminomas is similar to that identified in other solid malignancies [as reviewed in reference 45].

Embryonic stem (ES) cells are derived from the inner cell mass (ICM) of the blastocyst stage in early embryogenesis. These self-renewing cells progressively adapt to the culture conditions during continuous passages[14] and acquire an increased growth rate. Culture adapted ES cells are karyotypically abnormal, and the most frequently observed karyotypic changes are gains of materials from chromosomes 12, 17 and X, and similar non-random, recurrent alterations are seen in TGCTs as well [46]. Hence, the progressively culture adaptation of ES cells *in vitro* probably reflects malignant transformation *in vivo*.

EC and ES cells also show several other phenotypic resemblances, the most profound being pluripotency and the sharing of common cell surface markers and general gene expression programmes. Altogether, EC is commonly considered a malignant caricature of ES cells [47]. In this thesis, ES cells were considered a normal counterpart to EC cells in particular, and to TGCTs in general.

In this thesis, both early (karyotypically normal) and late passages (culture adapted and karyotypically abnormal) ES cell lines were included and compared to EC cell lines and TGCT tissue samples. ES cell line passages up to approximately 50 can be considered early, but this number is not well defined.

[14] Cell passaging – also known as sub-culturing, involves the splitting of cells and transfer of a smaller number to a new plate.

1.4 Colorectal cancer

World-wide, more than one million people are diagnosed with colorectal cancer (CRC) each year [48]. The corresponding number in Norway is about 3500, and CRC is the most common form of sex-independent cancer [49].

The majority of colorectal cancers occur sporadically, with less than five percent of the incidences being due to single hereditary components, as is the case for the autosomal dominant disorders Lynch syndrome and familial adenomatous polyposis (FAP). The relative life-time risk of developing CRC for Lynch syndrome and FAP individuals is 80 % and 100 %, respectively. Lynch syndrome, also known as hereditary non-polyposis colorectal cancer (HNPCC), is caused by a germ line mutation in one of the components of the DNA mismatch repair system (MMR), most commonly *MSH2* or *MLH1* [50]. FAP is linked to a germ line mutation in the tumour suppressor gene *APC*, which is involved in the degradation of β-catenin. Mutations leading to aberrant splicing of the genes associated with Lynch syndrome and FAP have been identified [51,52].

Colorectal cancer, sporadic as well as heritable cases, is considered to develop through different histopathological steps, resulting in either microsatellite or chromosome instable tumours (MSI and CIN, respectively) [53]. A subgroup of CRC with epigenetic instability, designated CpG island methylator phenotype (CIMP), has also been described [54]. CIMP overlaps with both MSI and CIN, but to a higher extent with the former. CIMP tumours are associated with promoter hypermethylation of a large number of genes, including *MLH1*, leading to sporadic MSI tumours. However, CIMP is characterised by distinct pathological, clinical, and molecular genetic features [55].

CIN tumours account for the majority (~ 85 %) of colorectal carcinomas, and are characterised by allelic losses, chromosomal amplifications and translocations [55]. Moreover, these tumours are most frequently found in the left side of the colon, and inactivation of the tumour suppressor genes *APC* and *TP53*, in addition to activation

of the proto-oncogene *KRAS*, are more prevalent among CIN tumours compared to MSI tumours [55].

Inactivation of the MMR system in sporadic MSI tumours is usually caused by epigenetic silencing of *MLH1*, leading to frameshift mutations and base-pair substitutions within randomly repeated nucleotide sequences known as microsatellites. The majority of microsatellites are located in non-coding areas; however, some tumour suppressor genes, *e. g. TGF-βRII* and *BAX*, harbour microsatellites in their coding sequences, leading to frequent frameshift mutations. Frameshift mutations of *CTNNB1* are also often seen in MSI tumours [55]. Additionally, MSI, and to a higher extent CIMP, is associated with an activating mutation of the proto-oncogene *BRAF* [56]. MSI tumours are generally located in the right colon, and MSI patients have a better prognosis than stage-matched CIN patients. Furthermore, MSI is associated with female gender and poor differentiation [55].

In 1932, Cuthbert E. Dukes proposed a classification system according to the extent of spread of the tumour [57] (Figure 7). However, the TNM staging system for colorectal cancer, first introduced by Pierre Denoix, is also commonly used. This describes three characteristics of the cancer; the extent of primary tumour (T) invasion, the degree of spread to regional lymph nodes (N), and whether or not metastasis (M) to distant organs is observed. Numbers (increasing with severity) are given for each of these characteristics, and then this information is combined.

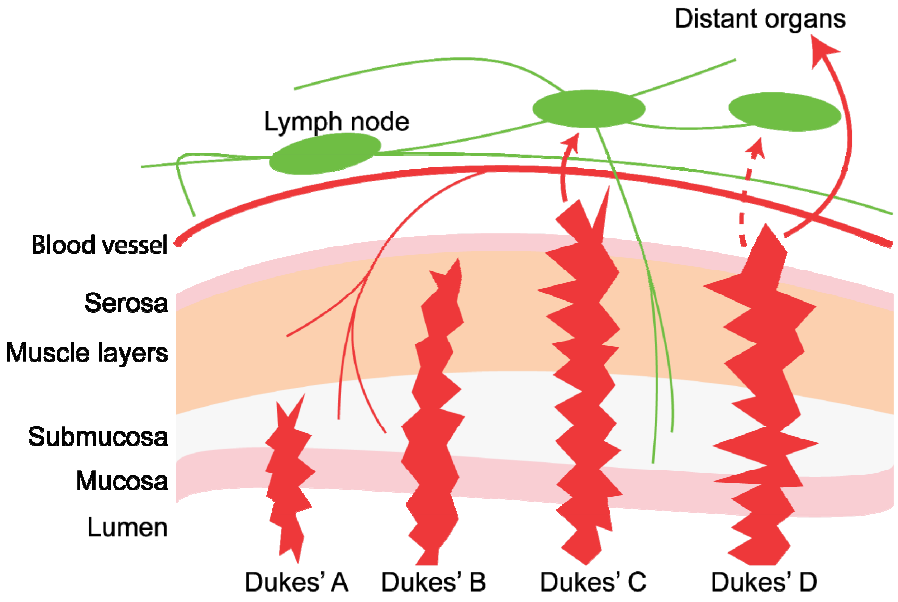


Figure 7. Modified Dukes' staging. In Dukes' A colorectal cancer, the tumour is confined to the mucosa and submucosa of the bowel wall, while in Dukes' B cancer, the tumour has penetrated into the muscle layer. In Dukes' C, cancer cells have spread to local lymph nodes, whereas in patients with Dukes' D, the cancer has metastasised to distant organs.

The prognosis among CRC patients depends to a great degree on the tumour stage at the time of diagnosis. Patients with a localised tumour have a 15-year survival close to 90 % (Dukes' A and B), and the corresponding number for patients diagnosed with regional spread is 65 % (Dukes' C). On the other hand, patients with distant metastases have a five-year survival of only about 10 % (Dukes' D) [49]. Hence, diagnosis at an early time point is important. However, there is no good screening strategy for colorectal cancer apart from colonoscopy. In this thesis, novel cancer specific transcripts were sought which may be implemented as diagnostic biomarkers for early detection of colorectal cancer.

2. Aims

The main aim of this thesis is to discover novel cancer-specific transcript variants which may serve as diagnostic biomarkers.

Specifically, we aim to establish an experimental pipeline to validate the existence of different transcript variants which were initially suggested by exon microarray data. Further, we intend to describe the biologically validated transcript variants in clinical samples from CRC and TGCT patients for presence of cancer-specificity, or whether there is a significantly difference in the ratios between alternative transcripts already present in healthy tissues.

3. Materials and methods

3.1 Materials

3.1.1 Cell lines

The project involved analyses of six colon carcinoma cell lines (HT29, HTC15, LS1034, RKO, SW48, and SW480), five EC cell lines (TERA1, TERA2, NTERA2, 2102Ep, and NCCIT), and the two ES cell lines H14 and Shef5, both early and late passages. After the exon arrays, no more RNA was available from Shef5 early passage, and consequently this sample was not included in the validation. Two EC cell lines, NTERA2 and 2102Ep, treated with retinoic acid to induce differentiation, were also included. RNA from all the colon carcinoma and ES cell lines, in addition to the samples from the three EC cell lines used to obtain exon array data, was isolated by the All prep DNA/RNA mini kit (Qiagen Co., Valencia, CA, USA). Trizol (Invitrogen, Carlsbad, CA, USA) was used in the isolation of RNA from the other cell lines. The end product of both methods is the total RNA fraction, but the column-based method do not preserve RNA fragments less than 200 nucleotides.

Culturing and sorting of the ES and EC cell lines were carried out in collaboration with the group of Peter Andrews at the Centre for Stem Cell Biology and Department of Biomedical Science, University of Sheffield.

3.1.2 Patient samples

Testicular germ cell tumours

Twenty-five TGCT samples, representing the five histologically different subtypes, were included in the project, in addition to six pre-malignant IGCN samples and five normal testicular parenchyma tissue samples. Frozen tissue sections stained with haematoxylin and eosin were evaluated by an expert pathologist to determine the

histological subtype of each of the TGCTs and IGCNs. For 23 of the samples, a frozen section of the exact same tissue piece as used in this study was evaluated, whereas for the remaining eight samples, the section was made from a neighbouring tissue piece. The surgical specimens were collected from several hospitals in Southern Norway during the years 1985 to 1992. RNA was isolated by Trizol (Invitrogen).

Colorectal cancer

One-hundred-and-seven primary colorectal carcinoma samples, of which 58 Dukes' B and 49 Dukes' C, were included in this study. Ninety-nine of these samples were collected at seven hospitals in the South-Eastern part of Norway between 1987 and 1989, and the remaining eight samples were collected at Aker University hospital in 2005 and 2006. From the same series as the eight last mentioned colorectal carcinoma samples, ten normal colorectal tissue samples were also included. After the exon arrays, there was no more RNA available from two normal and two cancer tissue samples. Consequently, these four samples were not included in the validation. RNA from cancer tissues was isolated by the All prep DNA/RNA mini kit (Qiagen), and total RNA from normal tissue samples was isolated by the RiboPure kit (Applied Biosystems/Ambion, Foster City, CA, USA).

3.2 Whole-transcript Expression Analysis

The processing of the exon microarrays and subsequent primary data analysis were performed by others in the research group. Descriptions of these analyses are yet included to facilitate an understanding of the completeness of the study.

In the TGCT study, exon arrays were run of RNA isolated from two ES cell lines (Shef5 and H14, both early and late passages), and three EC cell lines (NCCIT, NTERA2, and NCCIT). Exon arrays were also run of all the colon tissue samples mentioned in 3.1.2, except the cancer tissue samples collected at the Aker University Hospital in 2005 and 2006.

Total RNA was purified as described in section 3.1, and the RNA quality was assessed on Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). For each of the samples, one microgram of high quality RNA was used as input for the Affymetrix GeneChip Whole Transcript (WT) Sense Target Labeling Assay. Briefly, the samples were amplified and labelled according to the manufacturer's protocol. The resulting complementary DNA (cDNA) was then hybridised to the Affymetrix Human Exon 1.0 ST Array using the Hybridization Control Kit (Affymetrix) according to the manufacturer's protocol. Subsequently, the arrays were washed and stained using the Hybridization Wash and Stain Kit (Affymetrix) before scanning, all again according to the protocol.

The scanned array image was gridded by the Affymetrix GeneChip Operating System, which then calculated probe-level intensities from pixel values before storing these in CEL files. CEL data files were used as input for XRAY (Biotique Systems, Reno, NV, USA) to find significant differences in gene expression and transcript isoforms between the two groups. Quantile normalization and GC based background correction was applied, while p-values were corrected by using Benjamini and Hochberg False Discovery Rate (FDR).

In the TGCT study, exon-wise plots of the candidate genes with the most significant differences in exon usage were manually examined to generate a shortlist of genes for validation. The number of differential exon-level inclusions predicted by XRAY (Biotique Systems) in the colon cancer study were considered too numerous for manual inspection. Consequently, the FIRMA (Finding Isoforms using Robust Multichip Analysis) algorithm [58], found in the aroma.affymetrix BioConductor package, was utilized. In brief, the FIRMA scores each exon as to whether its probe signals deviate from the expected gene expression level. Accordingly, a difference in FIRMA score between two groups for a specific exon implies an event resulting in differential exon inclusion. The FIRMA score rankings were combined with XRAY p-values prior to manual accession of the candidates (Figure 8).

3.3 Validation of exon microarray data

Genes for which the initial microarray data analysis revealed a potential difference in the transcript structures were further evaluated both by use of *in silico* tools, as well as by wet-lab analyses (Figure 8). In the following, methods used to validate candidate cancer-specific transcripts will be described.

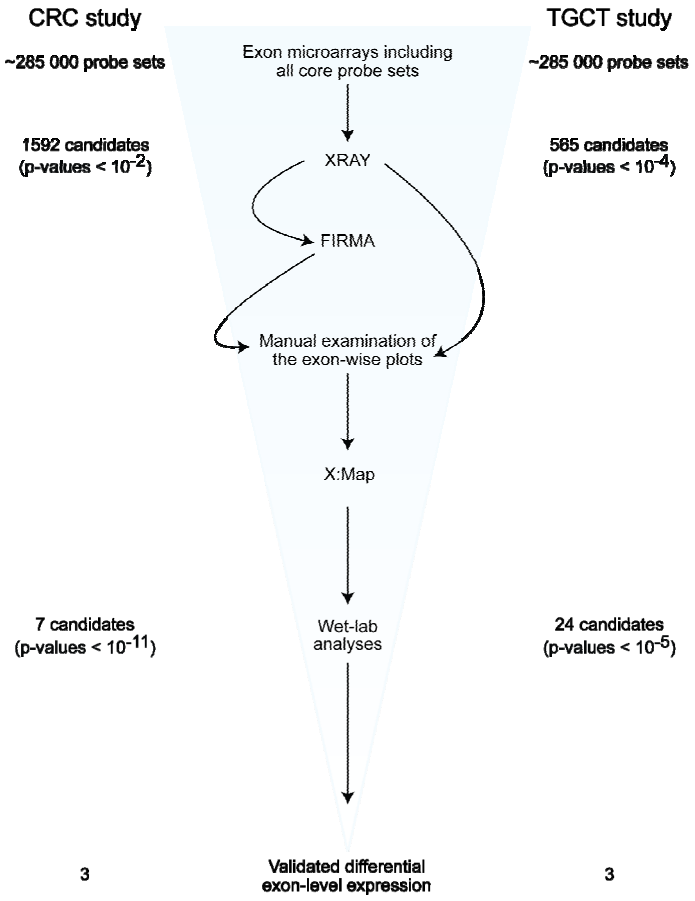


Figure 8. Filtering of candidate genes. In this figure, the CRC and TGCT studies are depicted to the left and right of the workflow, respectively. The starting point in both studies was the collection of all core probe sets, which refer to probe sets that are supported by the most reliable evidence from RefSeq and full-length mRNA GenBank records containing complete coding sequence information. Exon microarray data obtained from the two different groups were compared in the XRAY software (malignant versus normal tissues from colon and rectum, and ES versus EC cell lines), and significantly differential exon-level inclusions were listed. Then, the exon-wise plots were manually examined prior to visual inspection of the probe sets in the X:Map genome browser (mapping of probe sets to biological exons). In the CRC study, the XRAY p-values were combined with FIRMA score differences prior to examination of the plots. Subsequently, relevant assays were designed and the candidates were analysed in the laboratory. Finally, a total of six (three in each study) validated differential expression of transcript isoforms remained. Several candidate genes were rejected in each step. The master project involved examination of some of the exon-wise plots, in addition to all subsequent steps.

3.3.1 *In silico* exploration of candidate transcript variants

The interesting probe set(s) for each candidate gene, in addition to the surrounding probe sets, were searched for in the X:Map Genome Browser[15]. This was performed to map the PSRs to known biological exons and to visualise and inspect the candidate differentially included exon in annotated transcripts. The visual inspection of the probe sets in a genomic context is crucial with respect to designing a validation assay involving all transcript variants which potentially could be the cause of the divergent exon-wise profile, as the profiles do not reveal which exons are actually physically linked. Additionally, the X:Map Genome Browser gave warnings when searching for probe sets known to be cross-hybridising, and if this was the case for the probe set destined for validation, the respective candidate gene was rejected. Several other candidate differentially exon-level expressions were also rejected at this step, *e. g.* candidate events involving differential expression of an area in the middle of a known exon, as such events are relatively unlikely to be validated.

The visualisation of the probe sets in X:Map was coupled with the design of validation assays (RT-PCR primers), and probe sequences given in X:Map were searched for in the Ensembl Genome Browser[16]. Exon sequences, and in some cases intron sequences as well, were derived from Ensembl, and primers were designed.

3.3.2 Reverse transcriptase PCR (RT-PCR)

Primer design

In most cases, RT-PCR primers were designed to anneal to constitutive exons flanking the area of interest (Figure 9). This works well for simple cassette exons and alternative 5' and 3' splice sites, but other types of alternative splicing and transcription start/stop events are challenging and require more thoughtful primer

[15] http://xmap.picr.man.ac.uk

[16] http://www.ensembl.org

design, and this will be discussed in section 5.3. The forward and reverse primers were designed using the Primer 3 software[17] with default settings, and possible primer secondary structures and dimers were visualised and evaluated using the NetPrimer software[18]. All primers were purchased from MedProbe (Oslo, Norway). Primer information is available in Appendix I.

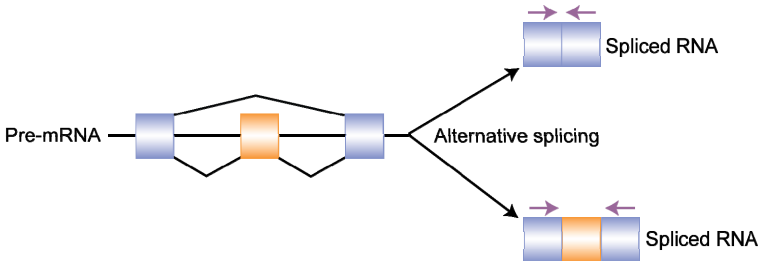


Figure 9. Primer design for RT-PCR validation of alternative splicing events. Primers are represented by purple arrows. The size of the RT-PCR product depends on whether the alternative exon, represented by an orange box, is included or skipped in the mature RNA.

Experimental assay

First-strand cDNA synthesis was performed using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). The RNA sample was diluted in RNase free water (Sigma-Aldrich, St. Louis, MO, USA) to an end concentration of 200 ng/µl in a total volume of 10 µl. Then 10 µl of the 2 x reverse transcription master mix containing RT buffer, dNTP mix, random primers, MultiScribe reverse transcriptase, and RNase free water (Sigma-Aldrich) was added. The reaction mix was incubated on a MJ Mini Gradient Thermal cycler (BIO-RAD, Hercules, CA, USA), first at 25 °C for ten minutes (primer annealing), then cDNA was allowed to synthesize at 37 °C for two hours, prior to reaction termination at 85 °C for five seconds. It is assumed that all the RNA is converted to cDNA in the reaction, hence the concentration of cDNA become 100 ng/µl. The cDNA was diluted to an end concentration of 50 ng/µl.

[17] http://frodo.wi.mit.edu

[18] http://www.premierbiosoft.com

The RT-PCR reaction was performed using the HotStarTaq DNA Polymerase Kit (Qiagen). One reaction mix contained 1 x PCR buffer, 200 µM of each of the dNTPs, 160 nM of each of the primers, one unit of HotStarTaq DNA Polymerase, and 50 ng of template cDNA. The PCR buffer includes $MgCl_2$ to a final concentration of 1.5 mM per reaction. If a concentration of Mg^{2+} of 3.0 mM was wanted, more $MgCl_2$ was added. The total volume was adjusted to 25 µl with Milli-Q water (Milli-Q Biocel System, Millipore Corporation, Billerica, MA, USA). The cDNA was amplified in a Robocycler Gradient 96 (Stratagene, La Jolla, California, USA), and the cycling conditions are shown in the table below (Table 1).

Table 1. Thermal cycling conditions RT-PCR.

			Additional comments
Initial activation step:	15 min	95 °C	Activation of the HotStarTaq DNA Polymerase
3-step cycling			
Denaturation:	30 sec	95 °C	
Annealing:	1 min	Variable	Depends on the T_m of primers
Extension:	2 min	72 °C	4 min for products > 2 kb (but < 4 kb)
Number of cycles:	27 or 30		Prior to fluorescent detection: 27 Prior to analysis on an agarose gel: 30
Final extension:	6 min	72 °C	

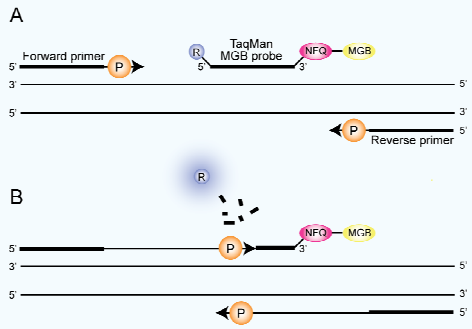
The number of cycles in the PCR reaction was determined by the method to be utilized to study the PCR products. Before capillary electrophoresis with fluorescent detection, the standard was to run 27 cycles, and prior to analysis on an agarose gel, the standard was 30 cycles. These fixed numbers of cycles were in some cases adjusted due to variation in the reaction efficiencies.

3.3.3 Real-time RT-PCR

Primer and probe design

For all the candidate genes destined for real-time analysis, except *ASPH,* there was no pre-designed TaqMan Gene Expression Assay (Applied Biosystems) available. For two of these, *ABCC3* and *PMF1,* the Primer Express Software v3.0 (Applied Biosystems) was used to design primers and probe. For the genes *SLC39A14, SPAG9* and *DNMT3B*, we were not able to modify the settings of this software to obtain a useful assay. In these cases, the Primer3 software was used with default settings to design primers, and the probe was manually designed. The probe was selected based on the parameters given by the Primer Express Software. The primer and probe melting temperatures and GC contents were analysed by the Primer Probe Test Tool found in the Primer Express Software, and if necessary, modifications were done. The commercial TaqMan minor groove binder (MGB) assay, in addition to the custom designed MGB probes, were purchased from Applied Biosystems, whereas the primers were purchased from MedProbe. The chemistry of the TaqMan assays is described in Box 1.

Box 1. Chemistry for the quantitative RT-PCR assays.



The TaqMan® MGB probe contains a reporter dye and a quencher dye linked to the 5' and 3' ends of the probe, respectively. When the probe is intact, the proximity of the reporter dye to the quencher results in suppression of the reporter fluorescence. The MGB raises the melting temperature (Tm) of the probe, which should be approximately ten degrees Celsius higher than the primer Tm to make sure that the probe hybridises specifically to its complementary sequence between the forward and reverse primer sites prior to primer annealing and extension. **(A)** The probe hybridises to the template, the primers anneal, and the polymerase starts adding deoxynucleotides. **(B)** When the polymerase reaches the probe, the 5'-3' exonuclease activity of the polymerase cleaves the probe, and the reporter dye is separated from the quencher, and the fluorescence from the reporter increases. The first cycle at which the fluorescence intensity is greater than the background fluorescence (a predefined threshold value), is denoted the Cycle threshold (Ct), and this value is representative of the starting copy number in the original template. Abbreviations: NFQ, nonfluorescent quencher; MGB, minor groove binder; R, reporter; P, hot-start DNA polymerase [modified from TaqMan® Gene Expression Assays Protocol PN 4333458 (Applied Biosystems)].

Experimental assay

The cDNA synthesis was performed using the same kit as previously described.

The pre-designed commercial quantitative RT-PCR assay was carried out in a fast optical 96-well reaction plate (Applied Biosystems), and the custom-designed assays were performed in standard 96- or 384-well optical reaction plates (Applied Biosystems). Different TaqMan master mixes, reaction volumes, and thermal cycling conditions were used with regard to whether the reactions should be carried out in fast or standard, or 384- or 96-well plates. The TaqMan Fast Universal PCR Master Mix (No AmpErase UNG, Applied Biosystems) was utilized in fast reactions, and the TaqMan Universal PCR Master Mix (AmpErase UNG, Applied Biosystems) was used in standard reactions. The final concentrations of master mix, forward and reverse primers, and probe in the standard reactions were 1 x, 0.9 µM of each, and 0.2 µM, respectively. In the fast reaction, the end concentrations of master mix and TaqMan Gene Expression Assay (primers and probe) were both 1 x. A total reaction volume of 20 µl was used when the reactions were performed in 384- and fast 96-well plates, as distinct from standard 96-well plates, where the total volume per reaction was set to 25 µl. RNase free water (Sigma-Aldrich) was added to the correct total volume. In each setup, the amount of starting material (cDNA) was 10 ng.

Multiplex real-time RT-PCR was done as well, adding one primer set, and two different probes with distinct dyes (FAM and VIC), in the same well. In these cases, the concentration of each of the primers where doubled compared to a standard setup, whereas the remainder (plate, master mix, final concentrations, volume, and thermal cycling conditions) was the same as for a standard assay.

The plates were incubated, and fluorescence measured, on an ABI 7900HT Fast Real-Time PCR System (also known as a “TaqMan”; Applied Biosystems). The thermal cycling conditions differed in the fast and standard reactions (Table 2).

Table 2. Thermal cycling conditions real-time RT-PCR.

		Standard	Fast
UNG activation:	50 °C	2 min	-
Polymerase activation:	95 °C	10 min	20 sec
Denaturation:	95 °C	15 sec	1 sec
Annealing and extension:	60 °C	1 min	20 sec
Number of cycles:		40	40

The pipetting robot EpMotion 5075 (Eppendorf, Hamburg, Germany) was used to pipette template to the wells in 384 plates, but the 96-well plates were set up manually. Master mix was distributed manually with a multi-channel pipette.

A standard curve was produced by serially diluted universal human reference (UHR) cDNA, synthesised from UHR RNA (Stratagene), of known concentrations. All samples were run in triplicates, and two endogenous control gene assays, *ACTB* and *GUSB* (both Applied Biosystems), were performed on all the samples in the TGCT study. In the colon cancer study, only the *ACTB* assay was utilised.

3.3.4 Detection of PCR products

Agarose gel electrophoresis

PCR products were separated according to size by agarose gel electrophoresis. The gels were made with 2 % agarose for separation of most products, but in some cases a 3 % gel was made to separate the bands more clearly. Eight or twelve grams of agarose (BioRad, Hercules, CA, US) were heated in 400 ml of 1 x Tris-acetate EDTA (TAE) buffer to make a two or three percent gel, respectively. The intercalating dye ethidium bromide (GeneChoice, Frederick, USA) was added to make the PCR products visible when illuminated with ultraviolet light. Ten to 25 µl of PCR product was mixed with the gel loading buffer bromophenol blue (one fifth of the volume of PCR product) and loaded onto the gel. The electrophoresis was run at 200 V for 25 to 30 minutes.

Capillary electrophoresis of fluorescently labelled PCR product

Prior to the capillary electrophoresis, a PCR was performed in which either the forward or reverse primer was labelled with the fluorophore 6-FAM. In each well of a 96-well optical reaction plate (Applied Biosystems), 0.5 µl of PCR product was mixed with 9.0 µl highly deionised (Hi-Di) formamide (Applied Biosystems) and 0.5 µl of a fluorescently labelled size standardised DNA ladder. The plate was then sealed with a 3100 Genetic Analyzer Plate Septa (Applied Biosystems), and the DNA was denatured for five minutes at 95 °C. Subsequently, the plate was placed in a 96-well plate base and inserted into a fully automated ABI 3730 DNA Analyzer (Applied Biosystems). Depending on the expected fragment sizes, either the ladder GeneScan 500 LIZ Size Standard or GeneScan 1200 LIZ Size Standard (Applied Biosystems) was chosen. The 500 LIZ was used with fragments up to 600 bp in length, and the LIZ 1200 was chosen when 600 to 1000 bp PCR products were expected.

The samples were analysed using the software GeneMapper v3.7 (Applied Biosystems). See section 3.3.5 for information on capillary electrophoresis.

3.3.5 DNA Sequencing

Purification of template DNA

PCR products separated on an agarose gel were cut out of the gel and purified using the MinElute Gel Extraction Kit (Qiagen) according to the manufacturer's protocol. DNA fragments purified with the MinElute system are ready for direct use in several applications, including ligation and transformation, and sequencing.

Sequencing reaction

Each sequencing reaction consisted of 1.0 µl of PCR product eluted from agarose gel, 1.5 µl of BigDye 5 x sequencing buffer (Applied Biosystems), 1.0 µl BigDye Terminator v3.1 premix (Applied Biosystems), 150 nM of forward or reverse primer, and Milli-Q water (Millipore) to a total volume of 10 µl. The reactions were

incubated at 96 °C for two minutes, followed by 25 thermal cycles of 15 seconds at 96 °C, five seconds at 50 °C, and four minutes at 60 °C. The thermal cycling was performed on a Robocycler Gradient 96 (Stratagene). The premix contains polymerase, dNTPs, and ddNTPs. The different ddNTPs are modified with distinct fluorescent labels.

Product purification

After the sequencing reaction, unincorporated dye terminators and other small molecules were removed. This was done by centrifuging the products through a Sephadex column, which separates molecules according to size. Small molecules will diffuse into the pores of the column, thus their passage through the column is delayed. Larger molecules, on the other hand, like the sequence products, will pass directly through the column. The column was made by loading Sephadex G-50 Superfine powder (GE Healthcare, Little Chalfont, Buckinghamshire, UK) on to a Multiscreen 96-Well Filtration Plate (Millipore). Three-hundred µl of Milli-Q water (Millipore) was added to each well, and the Sephadex was then allowed to swell for two hours, prior to centrifugation at 910 relative centrifugal force (rcf) for five minutes. To rinse the columns, 150 µl water was added to the wells before another round of centrifugation at 910 rcf for five minutes. Ten µl Milli-Q water (Millipore) and ten µl sequence reaction product was loaded on to each column, prior to centrifugation at 910 rcf for six minutes. The samples were collected in the wells of a 96-well optical reaction plate (Applied Biosystems), and the plate was then sealed with a 3100 Genetic Analyzer Plate Septa (Applied Biosystems). Subsequently, the plate was placed in a 96-well plate base and inserted into a fully automated ABI 3730 DNA Analyzer (Applied Biosystems).

Capillary analysis

In the ABI 3730 DNA Analyzer, the various fluorescent labelled fragments of different lengths are separated according to size in a 48-capillary array filled with POP7 polymer (Applied Biosystems). As the sequencing products pass the detection window, a laser beam excites the dye molecules and causes them to fluoresce. The

Data Collection software interprets the fluorescence data and displays them as electropherograms. The samples were analysed using the Sequencing Analysis 5.2 software (Applied Biosystems) and the free DNA sequencing chromatogram trace viewer FinchTV v1.4 (Geospiza, West Harrison, Seattle, USA).

4. Results

To identify novel transcript structures in cancer, as well as skewed ratios between pre-existing transcript variants, and evaluate their potential as cancer biomarkers, exon microarray data of two different cancer types were investigated. EC and ES cell lines were included to identify malignancy-specific transcript variants in a stem cell context (part 4.1), while malignant and normal tissues from colon and rectum were included to identify novel biomarkers for CRC (part 4.2).

4.1 Transcript variation in testicular germ cell tumour

The exon microarray data obtained from the ES and EC cell lines were analysed by others using the XRAY software for identification of differential expression at the gene and exon levels (only exon level analyses are discussed in this thesis). Five-hundred-and-sixty-five candidate events at the exon level had p-values of $< 1 \cdot 10^{-4}$, and were short-listed as candidates for malignancy-specific transcript variants. Subsequent manual examination of the exon-wise plots, and visual inspection of the probe sets in a genomic context, reduced the candidate list down to 24, which were selected for validation in the lab. Three of these candidates were confirmed, and are presented in the following (Table 3).

Table 3. Validated candidate genes in the TGCT study. Overview of the different transcript variants and the lengths of the PCR products representing them.

Gene	Transcript variant	Expression	PCR product
PMF1	Short-exon-four	High in NCCIT and 2102Ep	111 bp
	Long-exon-four	High in Shef5, H14 and NTERA2	172 bp
DNMT3B	Short	High in NCCIT and NTERA2	247 bp
	Long	High in Shef5, H14 and 2102Ep	436 bp
ZNF195	Short	High in EC (NCCIT, 2102Ep and NTERA2)	180 bp
	Long	High in ES (Shef5 and H14)	249 bp

4.1.1 Polyamine-modulated factor 1 (*PMF1*)

PMF1 is transcribed from the plus strand of cytogenetic band 1q22. The Ensembl Genome Browser[19], release 56, has annotated nine transcript variants of *PMF1*.

The exon microarray data are shown in Figure 10A. The exon-wise plot suggests that the overall expression of *PMF1* is somewhat higher in the EC cell lines 2102Ep and NCCIT compared to the ES cell lines Shef5 and H14. However, a probe set interrogating the 5' part of exon four shows the contrary regarding the expression, which implies differential splicing. However, notice that the EC cell line NTERA2 has an exon-wise profile resembling that of the ES cells. Two known splicing events, resulting in varying inclusion of sequence from exon four due to an alternative 3' splice site, were assumed to be responsible for this profile. Both standard and real-time RT-PCR assays were designed to validate the structure of the transcripts and their respective quantities (Figure 10B).

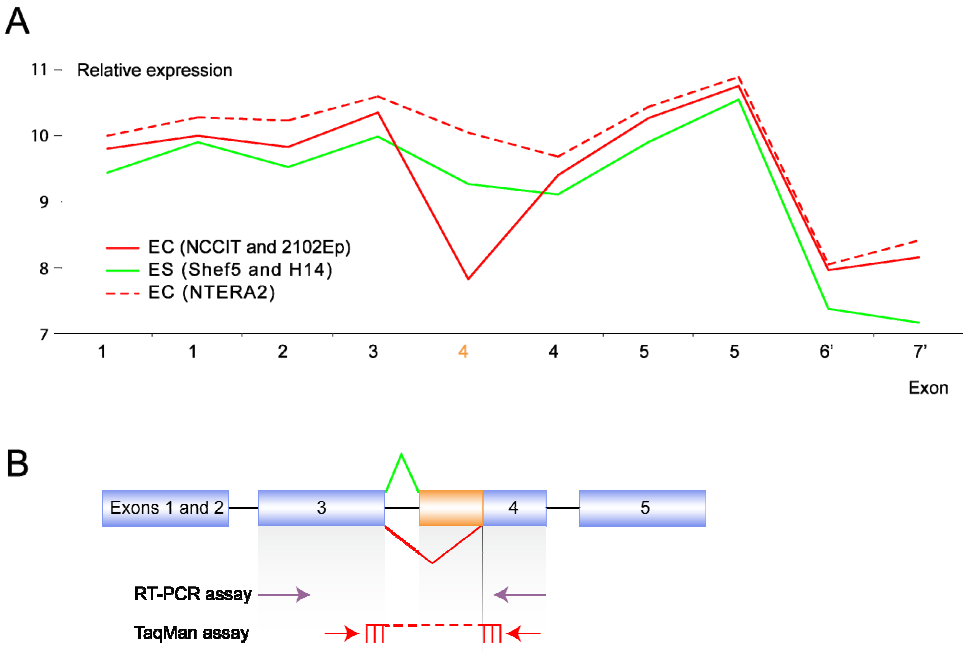


Figure 10. *PMF1* (ENSG00000160783). (A) Expression levels (log-2) of the different PSRs (often corresponding to the distinct exons) for *PMF1* as seen from exon microarray data.

[19] http://www.ensembl.org

The green and red lines represent the averages of four ES (Shef5 and H14, both early and late) and two EC cell lines (2102Ep and NCCIT), respectively. The EC cell line (NTERA2), which has an exon-wise profile resembling that of the ES cells, is plotted alone. Exons one to five are numbered according to the transcript ENST00000368277; however, exons 6' and 7' are not present in this transcript, but in other transcripts of this gene. **(B)** Two known splicing events assumed to be responsible for the interesting exon-wise plot are shown. The green and red lines represent the splicing events dominating in ES and EC cells, respectively. Primers for standard RT-PCR were designed to anneal to exon three and to the common part of exon four (purple arrows). The real-time RT-PCR (TaqMan) assay was designed as depicted, with a probe in the splice junction primarily present in the EC samples.

For *PMF1*, we chose standard RT-PCR and capillary electrophoresis of fluorescently labelled PCR products as a first-line validation, and included five of the same samples as examined by the exon microarray analysis (Figure 11). Here, we saw that the two ES cell lines (H14 and Shef5 late), and the EC cell line with an exon-wise profile resembling that of the ES cell lines (NTERA2), expressed the long-exon-four variant, resulting in a PCR product of 172 bp. However, the PCR product representing the short-exon-four transcript (111 bp) was barely visible. Conversely, for NCCIT and 2102Ep there was a shift towards the short-exon-four transcript, even though the long-exon-four transcript was still present at a level comparable to the other cell lines.

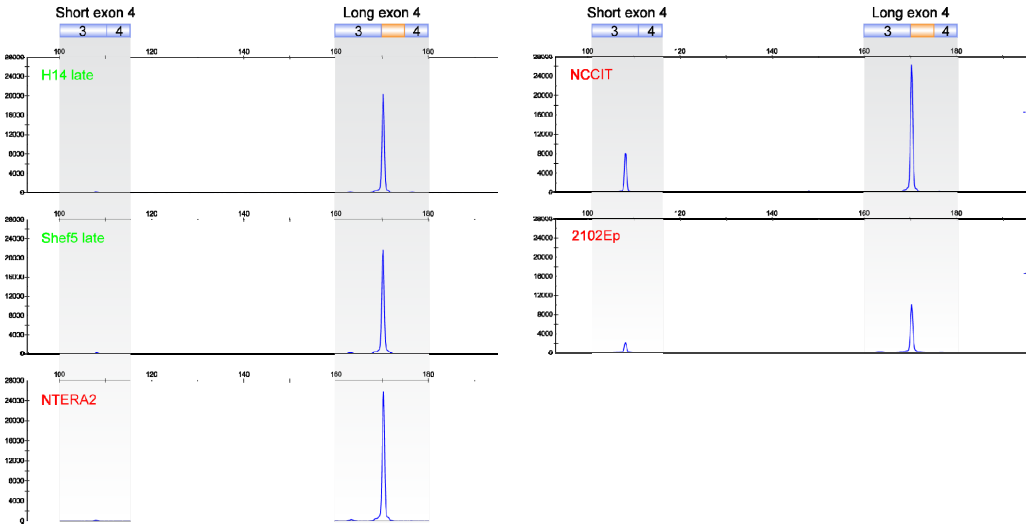


Figure 11. Validation of PMF1 transcript variants by capillary electrophoresis of fluorescently labelled PCR products. Depicted to the left are the three cell lines (H14, Shef5 and NTERA2) which based on the microarray data were expected to express relatively more of the long-exon-four, and less of the short-exon-four transcript, compared to NCCIT and 2102Ep (shown to the right). The graphs indicate the amount of RT-PCR product (y-axis; fluorescent intensity) relative to the RT-PCR product size in base pairs (x-axis). The lengths of the detected RT-PCR products were in accordance with the expected long and short-exon-four transcripts (172 bp and 111 bp, respectively).

On the basis of these results, we decided to investigate these transcripts further by expanding the sample series to include a series of clinical TGCT, IGCN and normal tissue samples, as well as ES and EC cell lines. In these samples, we analysed *PMF1* expression quantitatively, by real-time RT-PCR specifically targeting the short-exon-four transcript. The ratio between the amounts of this candidate cancer-specific transcript versus the long-exon-four transcript was determined by the RT-PCR assay already described. The short-exon-four transcript was detected by real-time RT-PCR in samples of several different histological subtypes. Interestingly, the samples separated clearly into two groups when these quantitative data were plotted against the transcript variant ratios (Figure 12). However, the groups were not separated by any obvious characteristics such as whether they originated from cancer versus normal samples, different histological subtypes (although all teratomas and seminomas, except one sample, are present in the group to the right), nor if they originated from tissue or cell line. Nevertheless, the samples were sorted along the x-axis as expected from the exon microarray data; the ES cell lines, in addition to the EC cell line NTERA2, form a part of the group down to the left, while 2102Ep and NCCIT are plotted to the right.

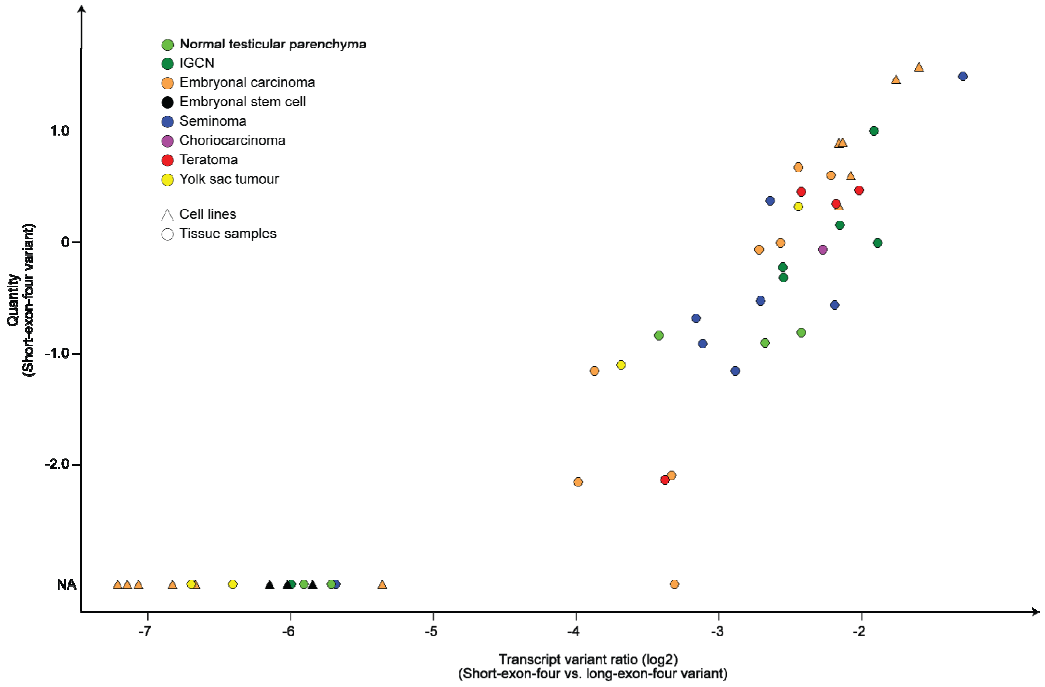


Figure 12. Quantitative and variant-specific *PMF1* expression data in a series of clinical TGCT, IGCN and normal tissue samples, in addition to ES and EC cell lines. Quantitative TaqMan data from the candidate cancer-specific transcript variant (short-exon-four transcript) are plotted against fluorescence intensity ratios (originating from capillary electrophoresis) between the short versus long-exon-four transcripts (both log-2). Several samples were not detected by the real-time RT-PCR, and these were denoted NA (not available).

4.1.2 DNA (cytosine-5)-methyltransferase 3β (*DNMT3B*)

DNMT3B is transcribed from the plus strand of cytogenetic band 20q11.21. The Ensembl Genome Browser, release 56, has annotated eight transcript variants of *DNMT3B*.

The exon microarray data are shown in Figure 13A. Two probe sets interrogating exons 20 and 21 proposes differential splicing, as exons 20 and 21 seem to have a relatively higher level of inclusion in ES cells, in addition to the EC cell line 2102Ep, compared to EC cells. Alternative inclusion or skipping of the two consecutive cassette exons 20 and 21 were suggested to be responsible for this profile. Both standard and real-time RT-PCR assays were designed to validate the structure of the transcripts and their respective quantities (Figure 13B).

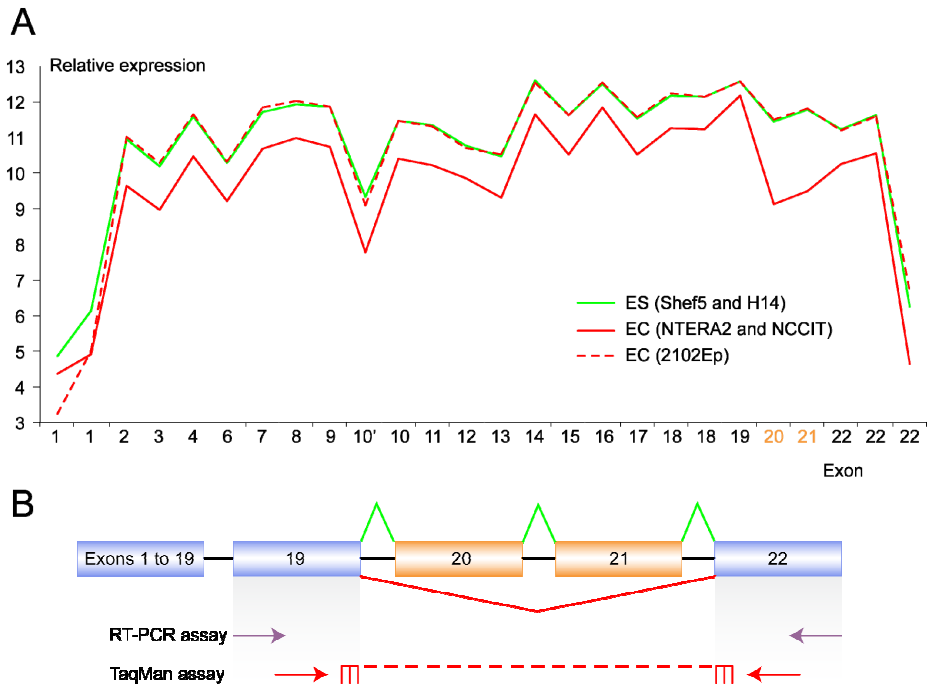


Figure 13. *DNMT3B* (ENSG00000088305). (A) Log-2 expression levels of the different PSRs for *DNMT3B* as seen from exon microarray data. The green and red lines represent the averages of four ES and two EC cell lines, respectively. The EC cell line (2102Ep) with an exon-wise profile resembling that of the ES is plotted alone. The exons are numbered according to ENST00000201963. **(B)** Two known splicing events assumed to be responsible for the interesting exon-wise plot are shown. The green and red lines represent the splicing events dominating in ES and EC, respectively. Standard RT-PCR primers were designed to anneal to the flanking constitutive exons 19 and 22. The real-time RT-PCR assay was designed with a probe in the splice junction primarily present in the EC samples.

In the first-line validation, standard RT-PCR and capillary electrophoresis were performed on five of the same samples as examined by the exon microarray analysis (Figure 14). Here, we saw that all the cell lines expressed both transcript variants. However, the two ES cell lines (H14 and Shef5 late), and the EC cell line with an exon-wise profile resembling that of the ES cells (2102Ep), expressed more of the exons 20 and 21 inclusion transcript variant (from now on denoted “long”, resulting in a PCR product of 436 bp) and less of the exons 20 and 21 skipping transcript variant (denoted “short”, resulting in a PCR product of 247 bp), while the opposite was true for the EC cell lines NCCIT and NTERA2.

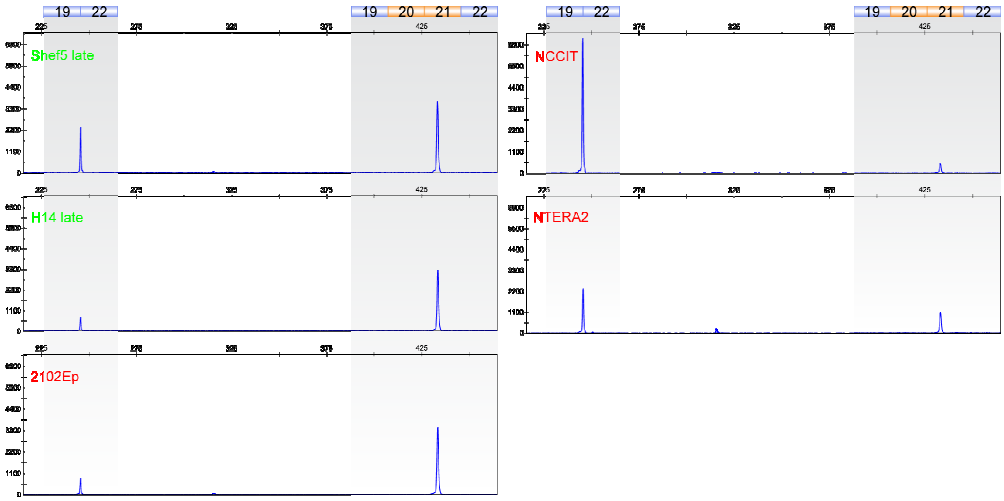


Figure 14. Validation of *DNMT3B* transcript variants by capillary electrophoresis of fluorescently labelled PCR products. Depicted to the left are the three cell lines (H14, Shef5 and 2102Ep) which based on the exon microarray data were expected to express relatively more of the long variant (exons 20 and 21 included), and less of the short variant (exons 20 and 21 skipped), compared to NCCIT and NTERA2 (shown to the right). The lengths of the detected RT-PCR products were in accordance with the expected long and short variants (436 bp and 247 bp, respectively).

The promising transcripts were further investigated by expanding the sample series to include a series of clinical TGCT, IGCN and normal tissue samples, as well as ES and EC cell lines. In these samples, we analysed *DNMT3B* expression quantitatively by real-time RT-PCR specifically targeting the short variant, and the ratio between the amounts of this candidate cancer-specific transcript versus the long variant was determined by the RT-PCR assay already described (Figure 15). The short variant was detected by real-time RT-PCR in all samples; however, the capillary electrophoresis detected only one of the variants in seven of the samples, and hence it was not possible to calculate any transcript variant ratio for these samples. This was also the case for an additional 20 samples, for which neither the long nor the short variant were detected by capillary electrophoresis. In general, the EC samples (cell lines as well as the tissue samples) and the ES cell lines showed a higher level of *DNMT3B* expression. Furthermore, the samples were sorted according to their transcript variant ratio as expected from the exon microarray data; the ES cell lines, in addition to the EC cell line 2102Ep (as well as the EC tissue samples), showed lower transcript variant ratio values than the EC cell lines NTERA2 and NCCIT.

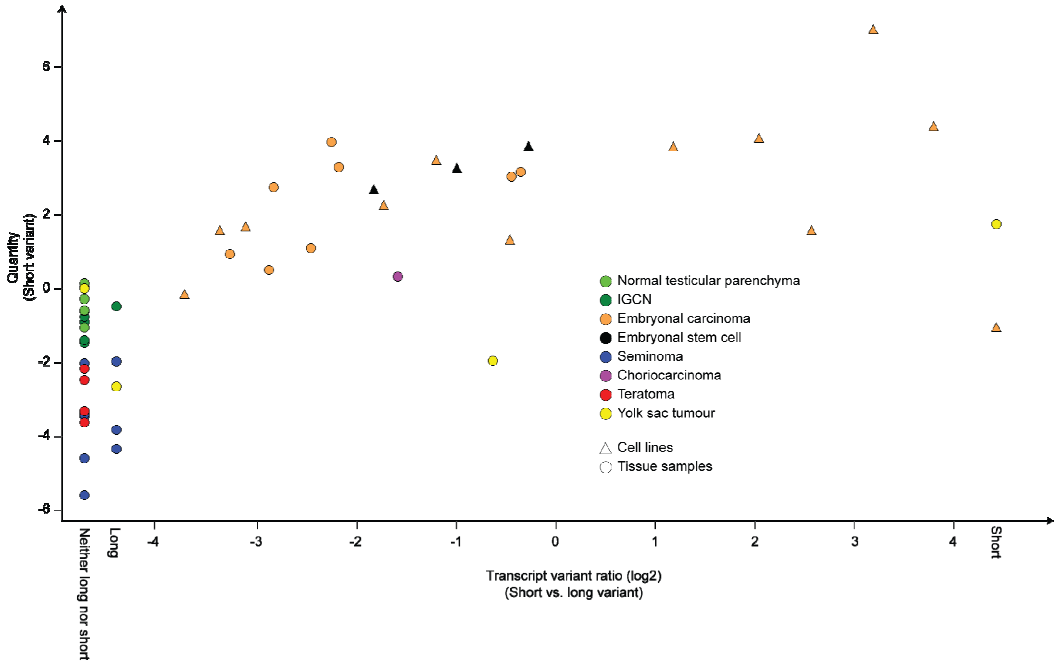


Figure 15. Quantitative and variant-specific *DNMT3B* expression data in a series of clinical TGCT, IGCN and normal tissue samples, in addition to ES and EC cell lines. Quantitative TaqMan data from the candidate cancer-specific transcript variant (short variant) are plotted against fluorescence intensity ratios between the short versus long variants from capillary electrophoresis (both log-2). The capillary electrophoresis detected only one of the PCR products in seven of the samples, either the long (five samples) or the short (two samples) product. Hence, transcript variant ratios could not be calculated for these samples. This was also the case for additional 20 samples (five of each of normal testicular parenchyma, IGCN and seminoma, in addition to four teratoma and one yolk sac tumour), for which neither the long nor the short variant was detected by capillary electrophoresis.

Additionally, both PCR products were sequenced which confirmed that the expected transcript variants were being analysed (Figure 16).

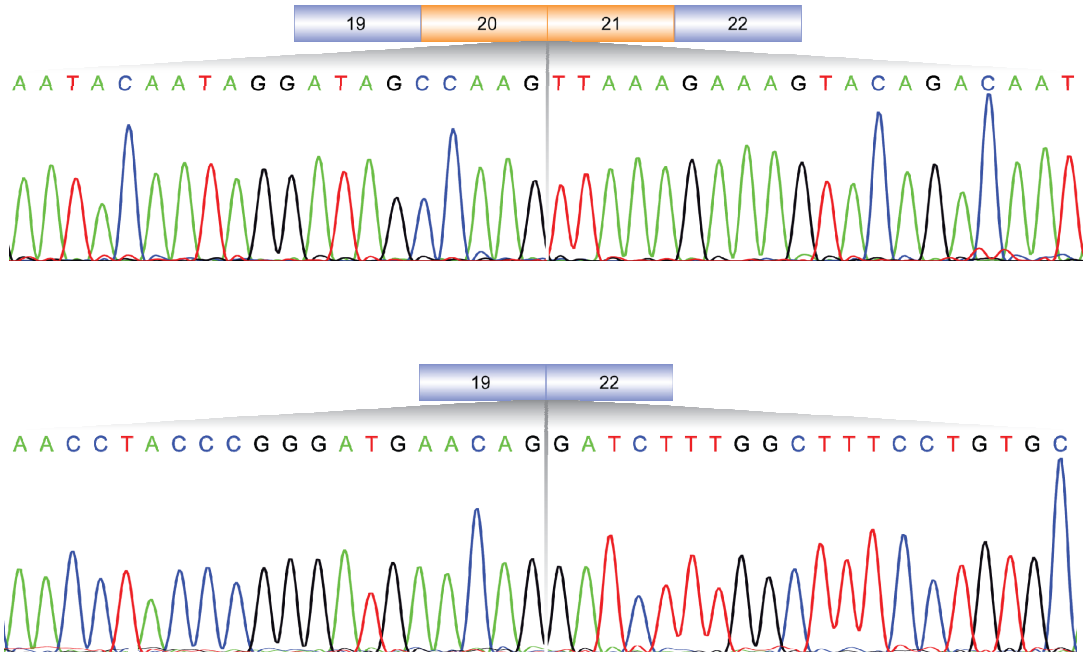


Figure 16. Sequencing of the *DNMT3B* RT-PCR products. Sequencing of the RT-PCR products confirmed that the expected transcript variants were amplified. On the top, 40 nucleotides in the junction between exons 20 and 21, only present in the long variant, are depicted; below, the same is the shown for the exon 19/22 junction belonging to the short variant.

4.1.3 Zinc finger protein 195 (*ZNF195*)

ZNF195 is transcribed from the minus strand of cytogenetic band 11p15.4. The Ensembl Genome Browser, release 56, has annotated seven transcript variants of *ZNF195*.

The exon microarray data are shown in Figure 17A. The probe sets interrogating exons two, four and six in *ZNF195* reveal similar expression of these exons in ES and EC cells. However, there are higher expression levels of the probe set targeting exon five in the ES cell lines. Two known splicing events, inclusion or skipping of the cassette exon five, were supposed to be responsible for this profile. A standard RT-PCR assay was designed (Figure 17B), and then performed on five of the same samples as analysed by the exon microarrays. By this method, all cell lines were found to express both variants, resulting in a long (exon five included, 249 bp) and a short (exon five excluded, 180 bp) PCR product. However, the ratios between the

amounts of short (high in EC) versus long (high in ES) PCR products were greater in the three EC cell lines as compared to the ES cell lines, thus reproducing and validating the exon microarray data.

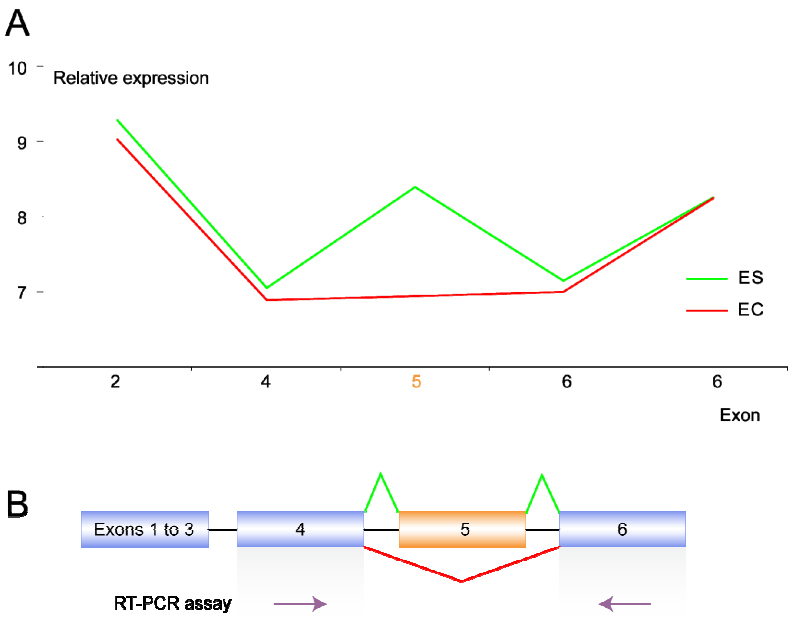


Figure 17. *ZNF195* (ENSG00000005801). (A) Expression levels of the different PSRs for *ZNF195* as seen from exon microarray data (log-2). The green and red lines represent the averages of four ES cell lines and three EC cell lines, respectively. Exons are numbered according to ENST00000005082. **(B)** Two known splicing events assumed to be responsible for the interesting exon-wise plot are shown. The green and red lines represent the splicing events dominating in ES and EC cells, respectively. Primers were designed to anneal to exons four and six flanking the alternative cassette exon five.

Subsequently, these transcripts were further investigated by performing the same standard RT-PCR assay as in the first-line validation on an expanded sample series including clinical TGCT, IGCN and normal tissue samples, as well as ES and EC cell lines (Figure 18). Plotting of the fluorescence intensity ratios (short versus long variant) against the mean peak heights (originating from the capillary electrophoresis), separated the samples into three groups; the ES cell lines, EC (both cell lines and tissue samples), and the other histological subgroups of TGCT, in addition to IGCN and the normal tissue samples. The ES cells and EC differed in the transcript variant ratios, while the other TGCT subgroups, in addition to the IGCN and normal tissue samples, differed in their overall expression of the gene.

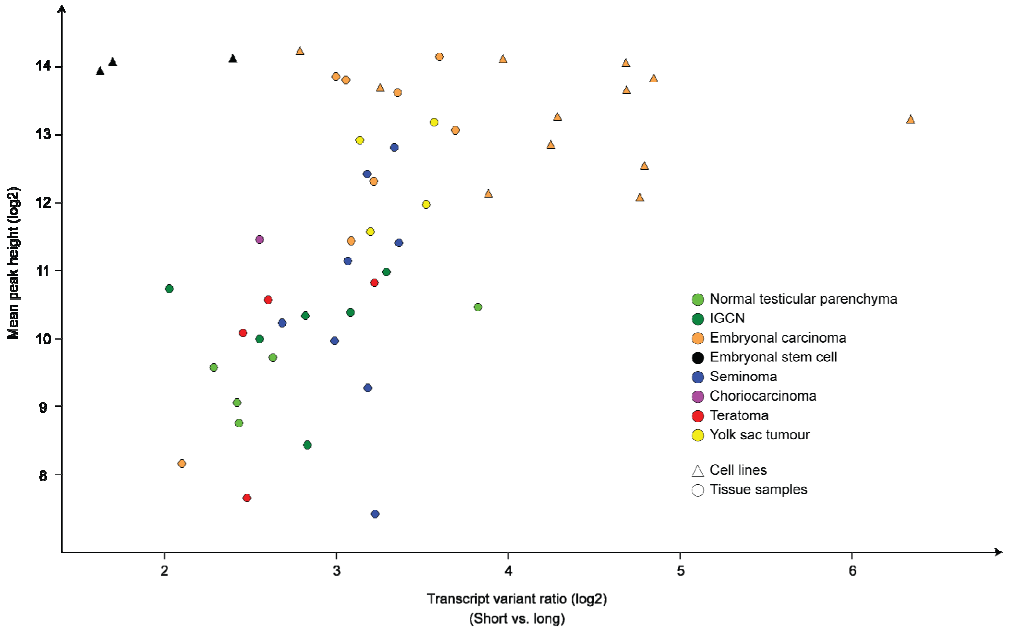


Figure 18. Differing *ZNF195* transcript variant ratio between EC and ES cells. Results from RT-PCR and capillary electrophoresis of fluorescently labelled *ZNF195* PCR products in a series of clinical TGCT, IGCN and normal tissue samples, in addition to ES and EC cell lines. The fluorescence intensity ratios between the short (candidate cancer-specific transcript) versus long variants are plotted against the mean peak heights (both log-2).

4.2 Transcript variation in colorectal cancer

Exon microarray data obtained from malignant and normal tissues from colon and rectum were analysed by others using the XRAY software for identification of differential expression at the gene and exon levels (only exon-level analysis are discussed in this thesis). The exon-level analyses yielded nearly 1600 candidate events, and the XRAY p-values were combined with FIRMA score differences between the two sample groups. Due to the great number of candidates, only a small portion of the top-scoring events were further investigated by manual examination of the exon-wise plots, followed by visual inspection of the probe sets in a genomic context. Seven genes were selected for validation in the lab, and three of these candidates were confirmed, and are presented in the following (Table 4).

Table 4. Validated candidate genes in the CRC study. Overview of the different transcript variants and the lengths of the PCR products representing them. For *SLC39A14*, no standard RT-PCR was performed.

Gene	Transcript variant	Expression	PCR product
SF1	Short-exon-two	High in tumour	266 bp
	Long-exon-two	High in normal	297 bp
DDX17	Short-exon-twelve	High in tumour	222 bp
	Long-exon-twelve	High in normal	603 bp
SLC39A14	Exon four	High in tumour	-
	Exon four primed	High in normal	-

4.2.1 Splicing factor 1 (*SF1*)

SF1, also known as branch point-binding protein (*BBP*), was considered a candidate gene in both studies; however, the most promising results were seen in the CRC study.

SF1 is transcribed from the minus strand of cytogenetic band 11q13.1. The Ensembl Genome Browser, release 56, has annotated 21 transcript variants of *SF1*.

The exon microarray data are shown in Figure 19A. The exon-wise plot implies that the overall expression of *SF1* is quite similar in malignant and normal tissues from colon and rectum. However, three probe sets interrogating a known extension of exon two resulting from an alternative polyadenylation site event, indicate differential inclusion of sequence from this area immediately downstream of the constitutive part of exon two. A multiplex RT-PCR assay[20] was designed (Figure 19B). In the first-line validation this assay was performed on six of the same samples as examined by the exon microarray analysis, three of each of malignant and normal tissues from colon and rectum, which were manually selected based on the exon-wise plots. By this multiplex RT-PCR assay, the exon microarray results were reproduced and validated.

[20] Multiplex PCR – variant of PCR which enables simultaneous amplification of several targets of interest in one reaction by using more than one pair of primers.

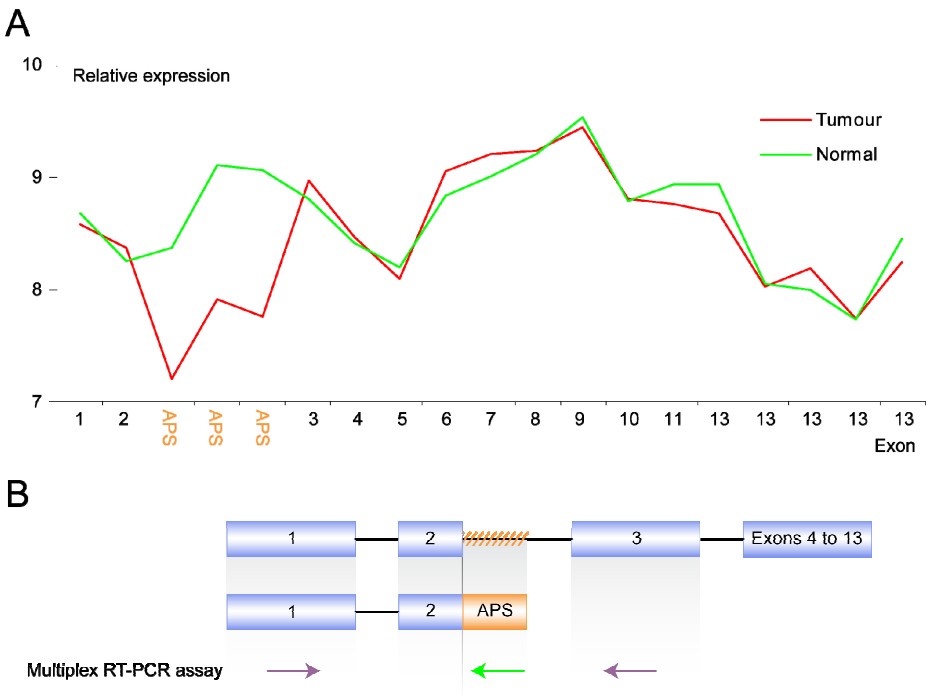


Figure 19. *SF1* (ENSG00000168066). (A) Expression levels of the different PSRs for *SF1* as seen from exon microarray data. The green and red lines represent the log-2 averages of the normal colon and CRC tissue samples, respectively. All exons are numbered according to ENST00000377390. APS (alternative polyadenylation site) represents an extension of exon two, and originates from an APS event. **(B)** The probe sets of interest, interrogating an intronic area close to exon two of the transcript used to name the exons (orange hatch), target the extension of exon two resulting from an alternative stop. Multiplex RT-PCR primers were designed to anneal to exons one, three and the extension of exon two as depicted. The reverse primer hybridising to the extension of exon two is coloured green, as it participates in the amplification of the transcript that seems to have a relatively higher level of expression in normal tissues, giving rise to a PCR product of 297 bp. The purple primers result in a PCR product of 266 bp.

Subsequently, these transcripts were further investigated by performing the same RT-PCR assay as used in the first-line validation on an extended sample series of clinical CRC and normal tissue samples, in addition to colon cancer cell lines. In short, the results show that the cancer tissue samples, in addition to the colon cancer cell lines, express a higher proportion of the transcript of the short-exon-two variant as compared to the normal tissue samples (Figure 20).

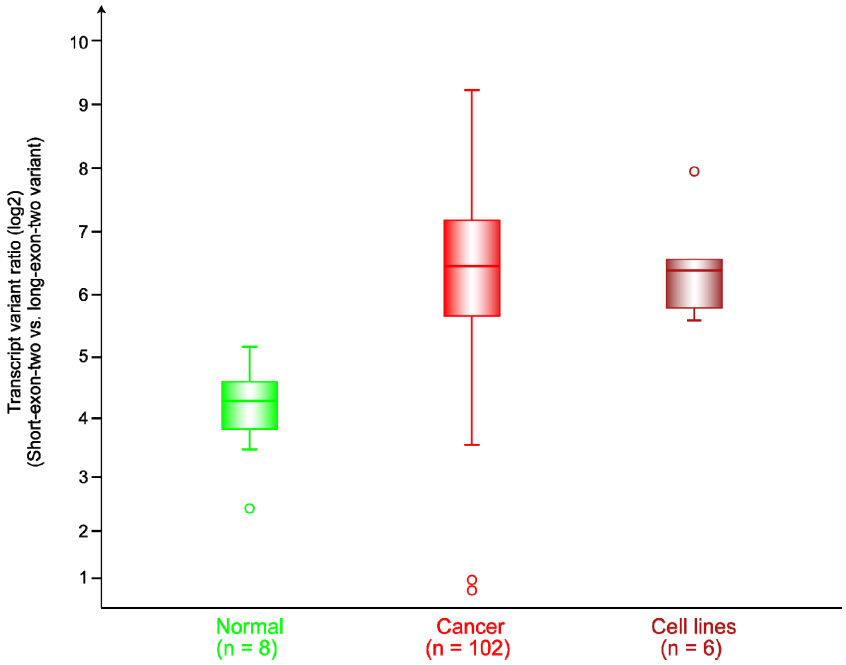


Figure 20. RT-PCR and capillary electrophoresis of fluorescently labelled *SF1* PCR products in a series of clinical CRC and normal tissue samples. The log-2 ratios between the amounts of short-exon-two (candidate cancer-specific variant) versus long-exon-two RT-PCR products are shown for malignant and normal tissues from colon and rectum, in addition to colon cancer cell lines. Three of the cancer tissue samples did only give rise to the short-exon-two RT-PCR product peak in the capillary electrophoresis, and consequently these are not included in the boxplot.

4.2.2 DEAD (Asp-Glu-Ala-Asp) box polypeptide 17 (*DDX17*)

DDX17 is transcribed from the minus strand of cytogenetic band 22q13.1. The Ensembl Genome Browser, release 56, has annotated 16 transcript variants of *DDX17*.

The exon microarray data are shown in Figure 21A. The exon-wise plot implies that a part of the intronic area between exons eleven and twelve (PSR mapping close to exon twelve in the genome), or possibly the whole intron (due to intron retention), has a higher level of inclusion in mRNA from normal tissue compared to CRC tissue. The first RT-PCR was performed with primers in exons eleven and twelve, but this yielded only one PCR product for every sample (size in accordance with expression of only exons eleven and twelve), which included six of the same samples as examined by the exon microarray analysis (three of each of malignant and normal tissues from colon and rectum). Based on this, a forward primer annealing to the

interesting area close to exon twelve was designed and run in multiplex with the former primer pair (Figure 21B), and for most samples two PCR products were synthesised. By this method, the exon microarray data were reproduced and validated. The interesting probe set was in a later release of the Ensembl genome browser found to target an extended exon twelve resulting from the use of an alternative promoter.

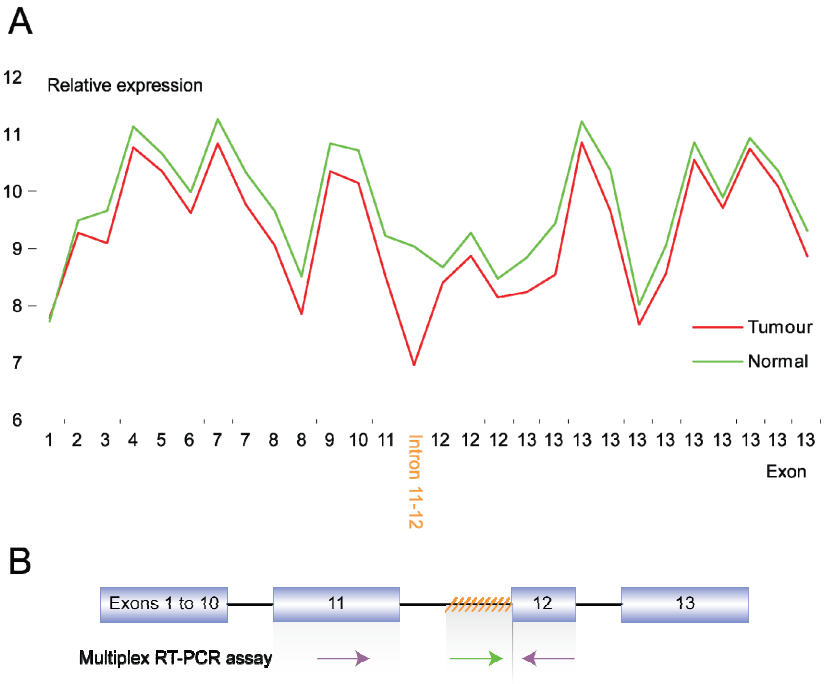


Figure 21. *DDX17* (ENSG00000100201). (A) Expression levels of the different PSRs for *DDX17* as seen from exon microarray data. The green and red lines represent the log-2 averages of the normal and CRC tissue samples, respectively. Exons are numbered according to ENST00000415456. **(B)** The probe set of interest, interrogating an intronic area close to exon twelve (orange hatch) was later found to target an alternative promoter (resulting in a long exon twelve). A multiplex RT-PCR assay was designed as depicted, resulting in two PCR products; the purple primer pair gives rise to a 222 bp PCR product, and primers internally in the long exon twelve give rise to a PCR product of 603 bp.

Subsequently, these transcripts were further investigated by performing the same multiplex RT-PCR assay as used in the first-line validation on an extended sample series of clinical CRC and normal tissue samples, in addition to colon cancer cell lines (Figure 22). In short, the results show that the CRC tissue samples, and to a lesser extent the colon cancer cell lines, have a higher proportion of the transcript of the short-exon-twelve variant as compared to the normal tissue samples.

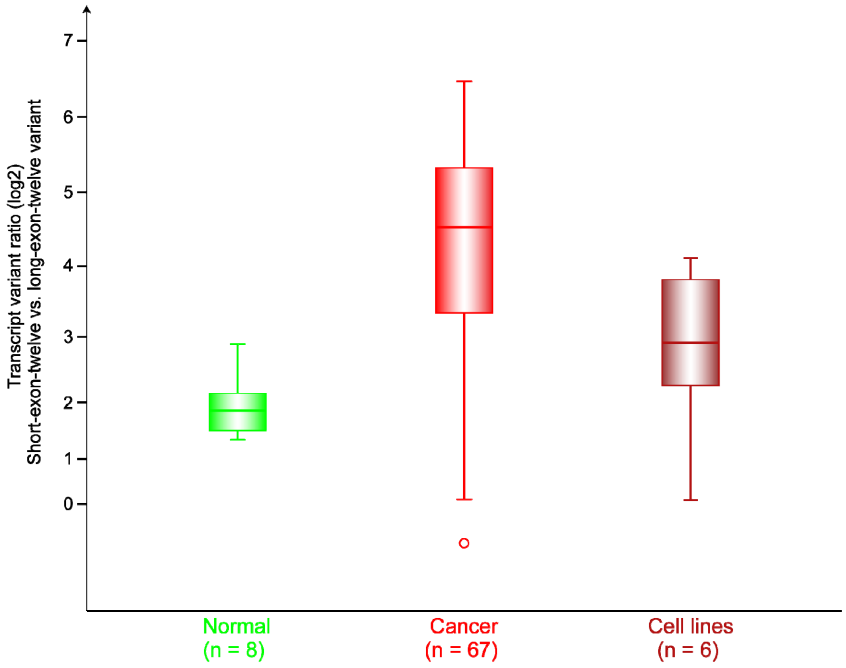


Figure 22. RT-PCR and capillary electrophoresis of fluorescently labelled *DDX17* PCR products in a series of clinical CRC and normal samples. The log-2 ratios between the amounts of short-exon-twelve (candidate cancer-specific variant) versus long-exon-twelve RT-PCR products are shown for malignant and normal tissues from colon and rectum, in addition to colon cancer cell lines. Thirty-eight of the 105 cancer tissue samples are not included in the boxplot, as 36 of them did only give rise to the short-exon-twelve RT-PCR product peak in the capillary electrophoresis, and two of them did not give rise to any peak.

4.2.3 Solute carrier family 39 (zinc transporter), member 14 (*SLC39A14*)

SLC39A14, also known as Zrt- and Irt-like protein 14 (*ZIP14*), is transcribed from the plus strand of cytogenetic band 8p21.3. The Ensembl Genome Browser, release 56, has annotated four transcript variants of *SLC39A14*.

The exon microarray data are shown in Figure 23A. From the exon-wise plot it seems that if exon four primed is expressed, exon four is not expressed, and *vice versa*, which implies that the two exons four are mutually exclusive. This assumption is reinforced by the fact that in known transcripts from this gene, the exons four-primed and four never exist in the same transcript. As the two exons four have identical lengths, it is impossible to differentiate between the two splicing events by standard RT-PCR with primers in flanking constitutive exons. Consequently, two real-time RT-PCR assays with exon specific probes were designed to validate the structure and

quantities of the transcripts resulting from these mutually exclusive splicing events (Figure 23B).

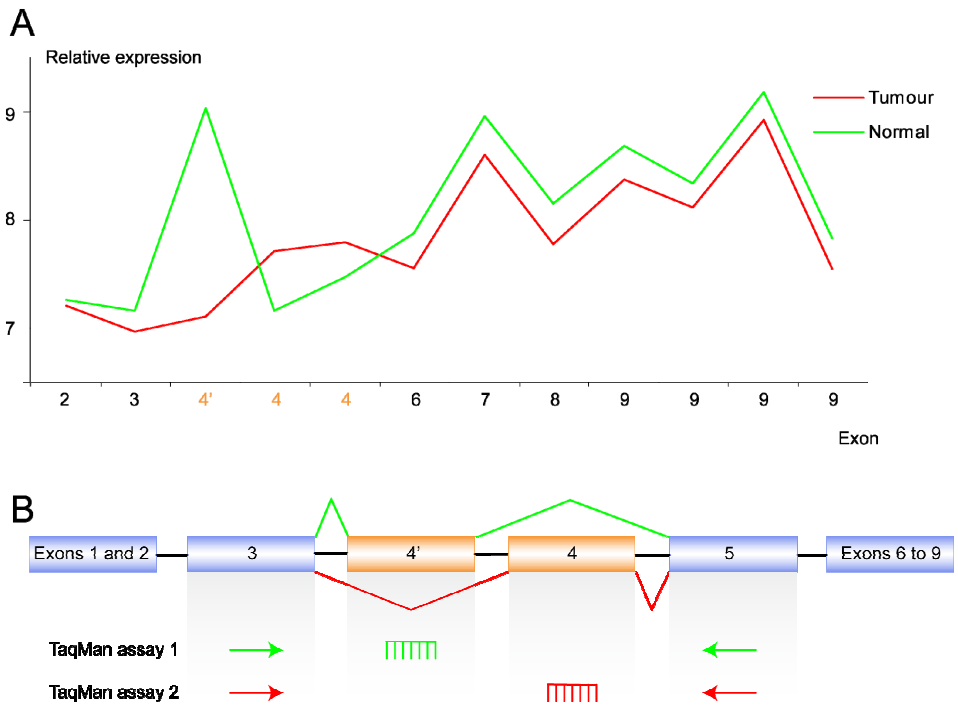


Figure 23. *SLC39A14* (ENSG00000104635). Expression levels of the different PSRs for *SLC39A14* as seen from exon microarray data. The green and red lines represent the log-2 averages of the normal and cancer tissue samples, respectively. Exons are numbered according to ENST00000289952; however, exon four-primed is not present in this transcript, but in another transcript of this gene. **(B)** Two known splicing events assumed to be responsible for the interesting exon-wise plot are depicted. The green and red lines represent the splicing events dominating in normal and cancer tissues, respectively. The two mutually exclusive exons four have identical size and similar sequences. Two real-time RT-PCR assays were designed with identical primers but distinct probes, as depicted.

In the first-line validation, these assays were performed on 19 of the same samples as examined by the exon microarray analysis, three normals (which were selected based on the exon-wise plots) and 16 CRC tissue samples, in addition to six colon cancer cell lines (Figure 24). This method revealed a clear difference between normal tissues and CRC samples (both tissues and cell lines). The real-time assay with a probe in exon four primed detected accumulation of PCR product in all three normal samples (all samples are run in triplicates), but only in two of 16 tumour samples and in none of the six cell lines. Furthermore, the real-time assay with a probe in exon four

detected PCR product in all cell lines, in addition to 15 of 16 tumour samples, but in none of the normal tissue samples.

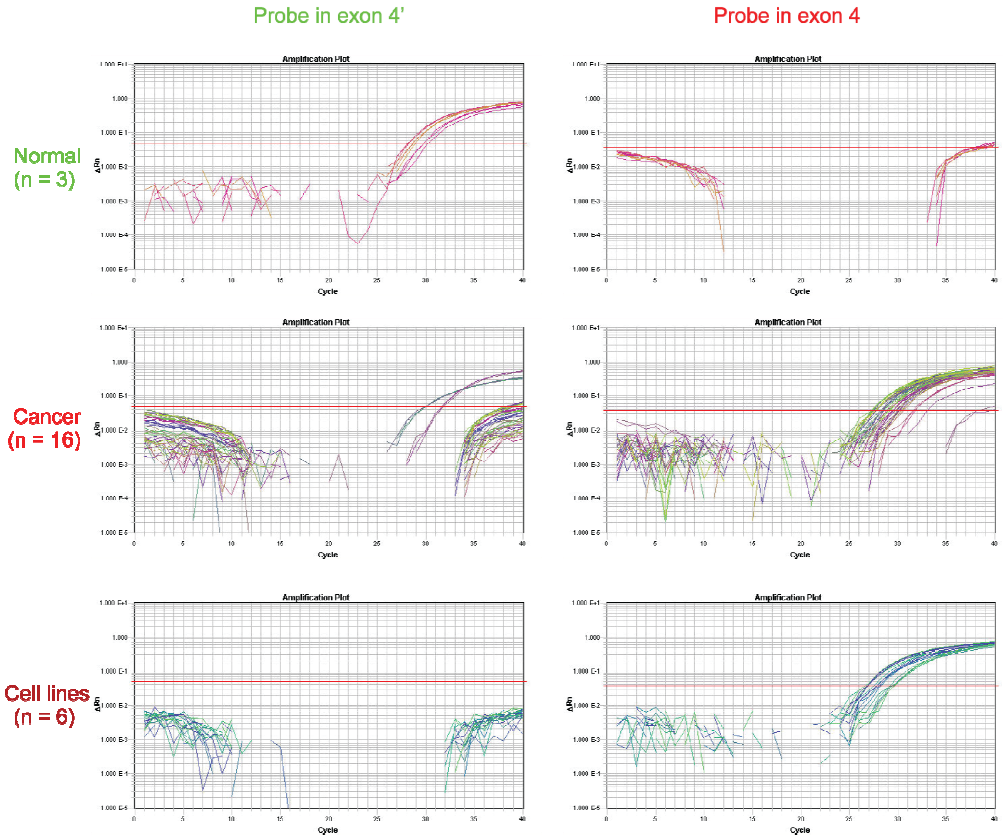


Figure 24. First-line validation of *SLC39A14* transcript variants by real-time RT-PCR. To the left, the real-time RT-PCR results (amplification plots) obtained with the assay containing the probe annealing to exon four-primed, which from the exon-wise plot seems to have a higher level of inclusion in normal tissue, are shown for all the samples included in the first-line validation; three normals, 16 CRC and six colon cancer cell lines. To the right, the corresponding data obtained with the probe hybridising to exon four are depicted. The fluorescence intensities (y-axis) are plotted against the number of PCR cycles (x-axis). The red horizontal line indicates the Cycle threshold (C_t).

Subsequently, these transcripts were further investigated by expanding the sample series of clinical CRC and normal tissue samples. The C_t values obtained for each of these samples by the assay with a probe in exon four-primed were normalised against the C_t values obtained with a probe in exon four, and the results are shown in Figure 25. Interestingly, the normal tissue samples consistently show negative relative expression values, and only two of 105 colorectal cancer tissue samples mix with the normal samples. Hence, setting a threshold at the highest value in the normal samples

yields a sensitivity of 98 % for this transcript variant. All the cell lines, and the great majority of the CRC tissue samples (97), show positive relative expression values.

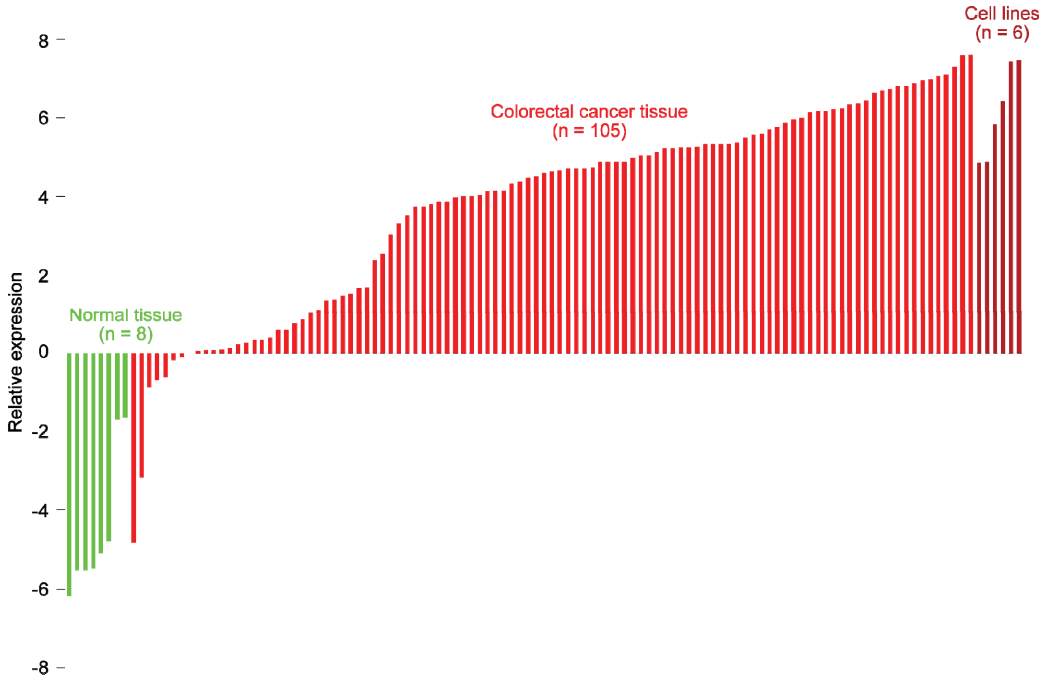


Figure 25. *SLC39A14* real-time RT-PCR data in a series of clinical CRC and normal tissue samples, in addition to six colon sell lines, showing the relative expression between the two mutually exclusive splicing variants. The C_t values obtained in the assay with a probe in exon four-primed ("normal exon") is normalised against the C_t values obtained in the assay with a probe in exon four ("cancer exon"), and this gives the relative expression along the y-axis (log-2) for each of the samples (x-axis). It is worth mentioning that all samples that did not cross the C_t line before cycle 34 were considered not to express the specific transcript, and were given the value 34 prior to the calculations.

5. Discussion

In the present project, the aim has been to identify specific transcript variants, known or novel, but all of which have resulted from exon-level transcriptome analyses of CRC and TGCT. In the bigger picture, transcript variants that are specific to cancer cells, or to particular subgroups of cancers, have good potential as biomarkers or even as drug targets.

The differential exon-level inclusion data destined for validation were generated using the Affymetrix Human Exon 1.0 ST Array, which contains probe sets interrogating more than one million annotated and predicted exons. With approximately four probes per exon, which means that for a gene containing ten exons there are roughly 40 probes matching its sequence, the gene-level estimates are generally in agreement with the true transcription level. However, the individual exon-level estimates are uncertain, leading to numerous false positive predictions of differential inclusion of sequence into mature RNA. Accordingly, it is decisive to investigate if the exon microarray data are reproducible by another method, *e. g.* RT-PCR. Biologically validated transcript variants were further examined to determine if the candidates were interesting in an extended clinical sample series.

Several of the candidate genes identified by exon microarrays and later passed validation in lab, have earlier been hypothesised to be implicated in cancer. Accordingly, it is not inconceivable that alternative splicing of these genes, or possibly alternative use of promoters and polyadenylation sites, might be functionally relevant in the development of cancer. Alternative RNA processing events may affect function by amongst others adding or deleting functional domains, changing affinities and altering mRNA stability.

In this study, cancer-associated transcript variants have been identified. In the following, a more detailed discussion of the results and the methods is provided.

5.1 Transcript variation in testicular germ cell tumour

Three alternative splicing events proposed by the exon microarray analysis comparing ES and EC cell lines were validated by RT-PCR. In all cases, the EC cell lines were found to have a relatively higher level of expression of a transcript missing a specific genomic area resulting from the use of an alternative 3' splice site (*PMF1*), or skipping of one or two consecutive cassette exons (*ZNF195* and *DNMT3B*, respectively).

DNMT3B, which encodes a *de novo* DNA methyltransferase, and *PMF1*, which encodes a part of the MIS12 kinetochore complex [59], and in addition is involved in polyamine homeostasis [60], have both been shown to have links to the development of cancer. For instance, *DNMT3B* overexpression induced by HOXB3 results in promoter hypermethylation and silencing of the tumour suppressor gene *RASSF1A* [61], and in another study, *PMF1* was identified to be epigenetically silenced in bladder cancer [62]. Polyamines are important regulators of proliferation and apoptosis [63], and defects in kinetochore formation may lead to aneuploidy and cancer; consequently, it is not inconceivable that *PMF1* silencing may be functionally relevant in the development of cancer. *ZNF195*, which may be involved in transcriptional regulation, might also have a role in cancer. This zinc finger protein, first identified in the human T lymphoma cell line Hut78 [64], was found to be selectively expressed by cutaneous T-cell lymphoma cells, and a clone expressing part of *ZNF195* was specifically recognized by sera of cutaneous T-cell lymphoma patients but not healthy donors [65]. Potentially, the alternative splicing of *DNMT3B*, *PMF1* and *ZNF195* could affect crucial protein domains and thus the activity and function of these proteins.

Real-time RT-PCR, in addition to standard RT-PCR followed by capillary electrophoresis, revealed that *DNMT3B* expression was higher in ES and EC (both cell lines and tissues) compared to the other histological subtypes of TGCT, in addition to IGCN and normal tissue samples. These results agree with what was found in a study examining and comparing gene expression in all the histological

subtypes of TGCT, which among other things revealed that high *DNMT3B* expression was significantly associated with EC [66]. The qualitative transcript differences between ES, in addition to 2102Ep, and NTERA2 and NCCIT revealed by the exon microarray analysis and then validated by RT-PCR, may only show that the EC cell lines NTERA2 and NCCIT differ from the other EC, as 2102Ep and the EC tissue samples resemble ES. If this is in fact the case, differential splicing of *DNMT3B* is less interesting in a cancer stem cell context based on these results.

Plotting of the real-time RT-PCR results against the transcript variant ratios for *PMF1* clearly separated the samples (normal testicular parenchyma, IGCN, TGCT and ES cell lines) into two groups. However, as mentioned in the results, the groups were not separated by any obvious characteristics. Accordingly, although the exon microarray data were reproduced within ES versus EC cells, the transcript variants of this gene seems to be of less cancer-specificity in an extended sample series.

Perhaps the most promising transcript variants identified in the TGCT study belong to the *ZNF195* gene. This gene is highly expressed in both ES and EC (cell lines as well as tissue samples); however, there is a clear difference in the short (candidate cancer-specific) versus long transcript variant ratios between the two different groups. The TGCT subgroups are mostly separated with respect to distinct combined expression of the two transcript variants, and the gene expression is generally lower compared to ES and EC. Altogether *ZNF195* is highly expressed in the pluripotency associated samples, and with a malignancy-specific expression pattern of the transcript variants.

5.2 Transcript variation in colorectal carcinoma

Three events resulting in differential expression of transcript variants proposed by the exon microarray analysis comparing malignant and normal tissues from colon and rectum were validated by other methods. By multiplex RT-PCR, normal tissues from colon were found to express a relatively higher level of transcript variants resulting

from the alternative use of a polyadenylation site (*SF1*) or promoter (*DDX17*) compared to malignant tissues and colon cancer cell lines. *SF1* and *DDX17* have both been shown to have links to the development of cancers. *SF1*, which binds the BPS of pre-mRNA (see section 1.2.1), has been suggested to have a role in colon tumourigenesis by being a component of the TCF-4 and β-catenin[21] complex and thus inhibiting the gene transactivational function of the complex [67]. *DDX17*, which encodes an RNA helicase amongst others involved in splicing, does also form a complex with β-catenin; however, the DDX17 protein (also known as p72) seems to promote the ability of β-catenin to activate gene transcription [68]. Potentially, the alternative use of a polyadenylation site (*SF1*) or a promoter (*DDX17*) could affect functional protein domains involved in gene regulation or changing the affinity for other transcription factors in the complex.

Two real-time RT-PCR assays targeting each of the mutually exclusive *SLC39A14* exons reproduced the exon microarray data, and this gene appears to be the most promising biomarker candidate revealed in this study. *SLC39A14* encodes a plasma membrane protein that functions as a zinc influx transporter [69]. Zinc is an essential cofactor for hundreds of enzymes [70], and this element is involved in the metabolism of macromolecules, as well as in the control of transcription, growth, development, and differentiation. The expression of one *SLC39A14* splicing variant instead of the other could potentially affect zinc homeostasis, and accordingly several cellular processes, including those relevant in the development of cancer. As early as in 1950, Addink proposed a correlation between zinc and malignant growth [71], and a study examining hepatocellular carcinomas found that *SLC39A14* is underexpressed in this form of cancer [72].

The two real-time RT-PCR assays detecting each of the *SLC39A14* transcript variants resulting from differential inclusion of two mutually exclusive exons differentiated

[21] β-catenin is involved in the out-migration of enterocytes from the colonic crypt. Accumulated cytosolic β-catenin diffuses to the nucleus and associates with other transcription factors and regulates expression of growth-promoting genes and genes involved in differentiation.

well between normal and malignant tissues from colon and rectum. For most samples this was an "either or" situation; six of eight normal tissue samples did exclusively express the transcript variant with exon four-primed ("normal exon"), whereas 78 of 105 (74 %) cancer tissue samples (and all the cell lines) did only express the exon-four variant ("cancer variant"). Including the remaining samples, only two of the 105 colorectal cancer tissue samples showed an expression ratio between the two different transcripts that mixed with the normal tissue samples. These real-time RT-PCR assays should be performed on an extended sample series including adenomas to determine if the shift in the expression of the transcript variants is an early event in the development of CRC. It is also worth mentioning that a standard curve for the real-time RT-PCR assay with a probe in exon four was produced by serially diluted UHR cDNA; however, this was not the case with the assay targeting the exon-four-primed variant as this was not detectable within the UHR reference sample. Accordingly, *SLC39A14* may have implication in other cancer types as well, as UHR cDNA contains cDNA from ten human cancer cell lines not including cell lines from colon and rectum.

5.3 Methodological considerations – Assay design

As mentioned in the description of RT-PCR primer design (section 3.3.2), validation of RNA processing events other than simple cassette exons and alternative 5' and 3' splice sites are often challenging. Some of these scenarios are discussed in the following.

RT-PCR and subsequent capillary electrophoresis of fluorescently labelled PCR products do not discriminate between two mutually exclusive exons of exactly the same size (Figure 26A). This situation may seem unlikely, but was the case for the two exons four in *SLC39A14*, which in addition had almost identical sequences. This challenge was solved by aligning the two exon sequences of interest and designing sequence specific TaqMan MGB probes targeting an area of difference, and real-time

RT-PCR was carried out with the same pair of primers annealing to the constitutive exons flanking the two exons four (Figure 26B).

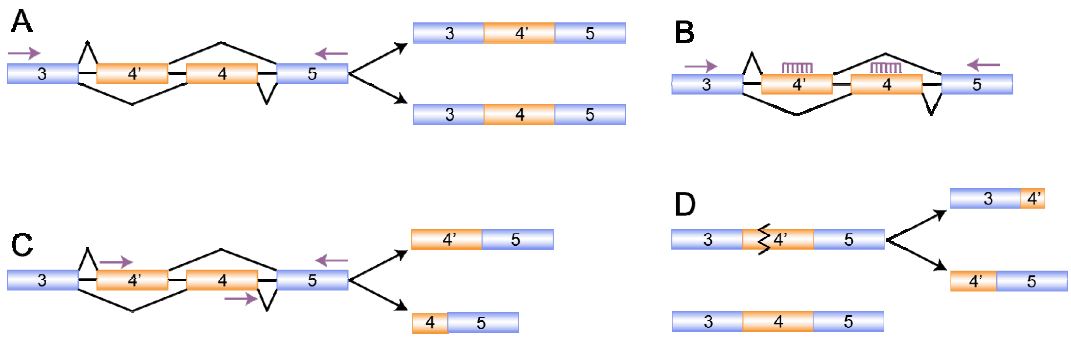


Figure 26. Examining splicing events involving mutually exclusive exons of same size. (A) Primers in constitutive exons flanking the two mutually exclusive exons of same size, resulting in PCR products of identical lengths, which are not separated by capillary electrophoresis. **(B)** In this thesis, this problem was solved by performing a real-time RT-PCR with sequence-specific probes targeting each of the two mutually exclusive exons four. Theoretically, it is possible to perform these PCR reactions in multiplex (the two different probes labelled by distinct dyes added in the same well); however, in this case only the single-plex setup (one probe in each well) worked. **(C)** In retrospect, it is evident that the two distinct transcripts could also have been distinguished by performing a multiplex PCR. **(D)** Additionally, the PCR products of different lengths in (A) could have been separated by capillary electrophoresis after treatment with a sequence specific endonuclease (represented by a "zig zag") that cuts only one of the exons four.

Prior to performing real-time RT-PCR, an alternative assay could have been considered (Figure 26C). Three primers could have been designed; one forward primer in each of the mutually exclusive exons, in addition to a reverse primer annealing to the downstream constitutive exon (or conversely, reverse primers in the mutually exclusive exons, and a forward primer in the constitutive upstream exon). The forward primers should have been placed in distinct relative positions within the two exons four, so that PCR products of separate lengths would have been synthesised representing the different transcript variants. This reasonable assay, with three primers in a multiplex PCR reaction, would have amplified the different transcript variants with distinct sets of primers, although the primer in the constitutive exon is mutual. For that reason, the PCR reaction efficiencies would have differed (*e. g.* due to distinct Tm optima of the two reactions). Hence, the true ratio between the different transcript variants could not have been deduced from the different amounts of amplified sequence from the two transcripts. Despite this inconvenience, this alternative assay could have been useful simply to examine if the gene was

interesting, as the amplification efficiency bias would have been the same for each sample. Therefore, if a difference between the two sample groups was observed, a follow-up real-time RT-PCR should have been performed. As the validation frequency of exon microarray data is rather low, it would normally not be profitable to perform a real-time RT-PCR on a candidate gene which has not passed a previous standard RT-PCR; however, in the case of *SLC39A14*, we were fortunate.

Even in cases where the same set of primers amplifies the distinct transcript variants, the PCR reaction efficiencies can differ due to different amplicon lengths, as shorter fragments are more efficiently synthesised. Nevertheless, this produces less bias than the multiplex scenario previously described.

A third option for distinguishing between transcripts with mutually exclusive exons of same size is to do a standard RT-PCR with primers in flanking constitutive exons, and then, prior to capillary electrophoresis, the PCR products can be cleaved by a restriction enzyme specific to one of the sequences (Figure 26D). However, in our example gene, finding a variant specific restriction enzyme is difficult due to the high degree of sequence homology between the two mutually exclusive exons.

Validation of alternative use of promoters and polyadenylation sites requires placement of a primer within the exon of interest because the target exon does not connect to flanking exons on both sides. In these cases, two separate PCR reactions should be performed; one that determines the presence or absence of the exon of interest, and another that measures the alternative pattern or the expression level of the gene. Accordingly, multiplex PCR reactions were carried out for *DDX17* and *SF1*, which represent alternative start and stop events, respectively. However, crucial *DDX17* and *SF1* transcript variants were not known at the time of the first primer design. Consequently, the initial primer pairs were designed to anneal to exons flanking the area of interest, and in each case these incomplete assays yielded only one PCR product (Figure 27).

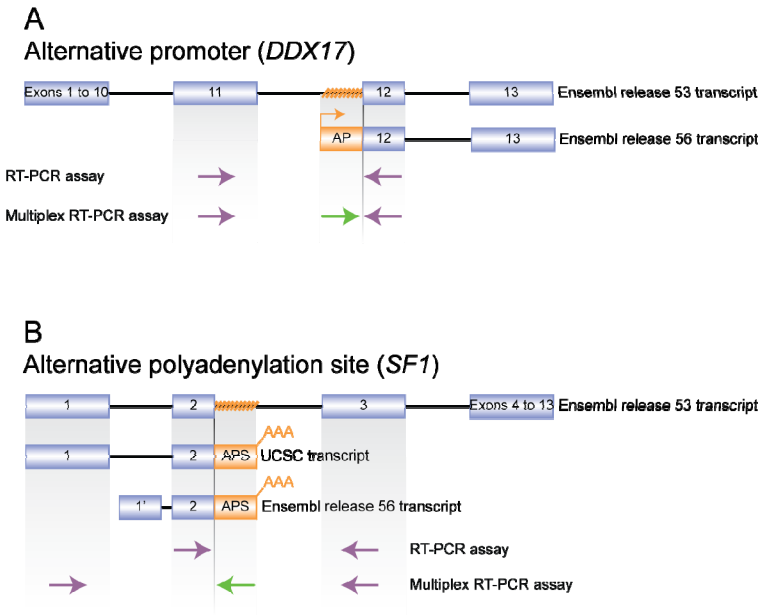


Figure 27. Validation of differential use of promoters and polyadenylation sites. In cases with alternative starts (here represented by *DDX17*) or stops (represented by *SF1*), a straight-forward strategy with a pair of primers in exons flanking the area of interest (orange hatch in Ensemble release 53 transcripts), as depicted in the RT-PCR assays in both **(A)** and **(B)**, leads to the synthesis of only one PCR product. These simple RT-PCR assays were designed due to ignorance of existing, but still unknown, transcript variants. The interesting probe sets in *DDX17* (targeting an area close to exon twelve) and *SF1* (close to exon two) targeted introns in the Ensembl release 53 transcripts, which aroused suspicion about unknown alternative 3' or 5' splice site events, respectively, or possibly intron retention. **(A)** As the straight-forward RT-PCR assay yielded only one product, a forward primer targeting the area of still unknown nature, appearing from the exon-wise plot to have a higher level of inclusion in normal tissue as compared to CRC tissue, was designed and run in multiplex with the primer pair constituting the previous RT-PCR assay. The ratio between the two synthesised PCR products differed between normal and CRC. Subsequently, Ensembl release 56 displayed an alternative promoter (AP) in the interesting area of *DDX17*. **(B)** An interesting transcript from the database of the University of California Santa Cruz (UCSC)[22] with an extended exon two resulting from alternative polyadenylation site (APS) usage was overlooked, and the RT-PCR assay design was accordingly solely based on known transcript variants in Ensemble release 53 and the exon-wise plot, and this assay yielded only one PCR product. Subsequently, when the nature of the interesting area was noticed, a multiplex RT-PCR assay with one primer in each of the exons one, three and the extension of exon two was designed as depicted. Later, one more interesting transcript containing the extended exon two was found in Ensembl release 56. This transcript did not express the part of exon one targeted by the forward multiplex primer. Consequently, if this transcript also contributes to the divergent exon-wise profile, the performed assay is incomplete. This may be the case in the TGCT study, as this gene was a candidate in both studies, but it was only validated in the CRC study.

These scenarios illustrate the importance of manual inspection and visualisation of the exon microarray data in a genomic context prior to designing assays for

[22] http://genome.ucsc.edu

validation, as exon microarrays do not reveal which exons are actually physically linked. However, incomplete collections of transcript variants in the genome browsers may lead to the design of assays not including all the transcripts responsible for creating the divergent exon-wise plots. During the master project, the number of known transcript variants in the public data bases has virtually exploded, reflecting the true complexity of biology. For instance, Ensembl release 52 (December 2008) has two transcript variants annotated for the *ING4* gene, whereas Ensembl release 56 (September 2009) alleges that *ING4* has a total of 23 transcripts. Additionally, several transcript variants have been withdrawn. *GCNT3* represents a frightful instance of this; both Ensembl release 53 and 56 revealed two *GCNT3* transcripts, apparently the same, as the transcript numbers were matching; however, the ENST00000267857 transcript contains different exons in the different Ensembl releases. If the last Ensembl release represents the true biology of this gene, the RT-PCR validation assay, which was based on Ensembl release 53, is deficient. However, with such wide-spread changes in the annotation during a 10-month period, it is likely that the current information in the Ensembl database is still quite far from the real-life situation. Thus, detailed transcript validation analyses of genes and probesets resulting from exon microarray analyses would have been warranted even if the technology itself were producing perfect data. Data generated from deep sequencing of cDNA fragments have barely started to accumulate; hence, it is expected that the complexity of the known transcriptome will dramatically increase.

Finally, it should be noted that if an intron retention event was expected of a large intron, the elongation time in the RT-PCR thermal cycling was increased. Subsequently, the RT-PCR products were analysed by capillary electrophoresis as well as on an agarose gel, as the former detection method is somewhat difficult to perform on larger fragments.

5.4 Methodological considerations – Detection of PCR products

To differentiate between inclusion and skipping events on an agarose gel, the overall size of the amplicon should be designed relative to the size of the interesting area. The smaller the interesting area, the smaller the amplicon sizes should be to clearly separate the skip/include products on a gel. In some cases it was difficult to design assays yielding PCR products that could be resolved on an agarose gel, and in other cases, the differences in band densities were not easily evaluated. Therefore, several of the genes were analysed by capillary electrophoresis and detection of fluorescently labelled PCR products. During the course of the project, it was decided to apply this method as the default strategy, even if a gel-based assay could have given the desired results. The capillary electrophoresis enabled improved resolution of PCR products of approximately the same size (generally down to one nucleotide in size difference). In addition, this fluorescence-based method is much more sensitive and also produces fairly quantitative results, thus enabling semi-quantitative measures of the ratio between different transcript variants. One reason for being more quantitative is that the high sensitivity allows termination of the PCR reaction at an earlier cycle and thus measurements of accumulated product in a phase more representative of the transcript input amount. However, real-time RT-PCR, which reveals the amplification plot, is the superior method when more accurate quantitative measurements are required. Theoretically, there is a quantitative relationship between the amount of starting template and the amount of PCR product at any given cycle. In real-time PCR, the quantity of PCR product is measured after each cycle, as opposed to traditional PCR, where the accumulated target sequence is analysed only after a fixed number of cycles.

In this thesis, sequence specific TaqMan MGB probes were utilized in the real-time RT-PCR experiments, and consequently only the synthesis of target DNA is measured. This method is costly and challenging compared to standard PCR, and was therefore reserved promising candidates, most of which had passed a previous standard RT-PCR. However, in most situations it was difficult to design a real-time

assay targeting both transcript events, and hence the real-time RT-PCR was performed to determine the quantity of one of the transcripts, and standard RT-PCR followed by capillary electrophoresis were carried out to obtain the ratio between the amounts of the different transcript variants. Multiplexing of two real-time RT-PCR reactions, which involve amplification of both transcript variants with the same set of primers, and detection of the different transcripts by distinct probes (labelled with different reporter dyes, *e. g.* FAM and VIC) added in the same well, would provide the most accurate estimate on the ratio between the amounts of the different transcript variants. However, the attempt to do this for the *SLC39A14* gene was unsuccessful, perhaps due to oligo dimerisation. Additionally, multiplex RT-PCR was planned for the *DNMT3B* gene, yet this was impossible as the long PCR product (295 bp) was not amplified by the TaqMan thermal cycling (the real-time RT-PCR product was analysed on an agarose gel), neither in single-plex nor multi-plex, maybe owing to a too short elongation time. Nevertheless, Applied Biosystems informs that the enzyme can synthesise 2-4 Kb/min, and consequently this should not be a problem, as the combined annealing and elongation time was one minute.

5.5 Validation frequency

For the candidate genes tested in the lab, the validation rate in the CRC and TGCT studies were 43 % (3 out of 7) and 13 % (3 out of 24), respectively, whereas the total validation frequency was 20 % (6 out of 30; *SF1* was a candidate gene in both studies, but only validated in the CRC series).

The validation rate in the CRC study was quite good compared to other studies qualitatively examining colon cancer transcriptomes, which confirmed 26 % [73] and approximately one-third [35] of the differential splicing events proposed by exon microarrays. The overall validation rate in this thesis was reduced by the candidates resulting from the exon microarrays comparing ES and EC cells. This is probably due to the fact that these two sample types are very similar in most respects, including at the transcriptome level. Consequently, the candidate event p-values destined for

examination in the lab in the CRC study were in general lower than in the TGCT study. Furthermore, the FIRMA algorithm (section 3.2) was only utilised in the CRC study. The use of this algorithm, followed by manual examination of the exon-wise plots, may have revealed a greater proportion of the false positives than by manual examination alone, which was the case in the TGCT study.

The design of an incomplete RT-PCR assay not including all the transcript variants responsible for the divergent exon-wise profile, may lead to rejection of a true positive candidate gene, which accordingly is wrongly categorised as a false positive (discussed in section 5.3). However, numerous actual false positives are generated by the exon microarrays and contribute considerably to the rather low validation frequency. In the following, scenarios that might result in such artificial predictions of qualitative transcriptome differences will be discussed.

As mentioned in section 1.2.3 (Global analysis of transcript variation), DNA microarrays have traditionally interrogated only the 3' ends of RNAs. For oligo DNA microarray providers other than Affymetrix, generally one individual probe, typically of about 60 nucleotides in length, is used per gene. The Affymetrix arrays use 25mer oligo probes, and the 3' biased microarrays from this vendor has used a collection of 11 probes, constituting a probe set, for each measurement. For the Affymetrix exon microarrays, there are individual probe sets targeting each individual exon, but each probe set is here reduced to include four probes, which again results in less reliable values for the individual probe sets, and hence a greater portion of false positives.

Furthermore, using exon-level detection instead of gene-level detection forces the issue of multiple comparisons, where the larger the number of exons in the gene, the greater the chance of high statistical significance by chance alone. This problem should be corrected for by the use of Benjamini and Hochberg FDR; however, several false positives arising from multiple comparisons will not be filtered out.

The generation of background noise in the exon microarray data, *e. g.* due to low sample quality, may also contribute to false positives. This is of great importance to

candidates that have large gene expression differences between the two groups, as a large disparity in transcription rate makes it more likely that probe set intensities in the two groups are disproportionately affected by the noise.

The expression level of a given exon is a result of the transcription rate of the entire gene and the inclusion rate of that exon. Therefore, to reveal differential inclusion of a specific PSR in mature RNA, it is crucial to normalise the probe set intensity to the expression level of the gene harbouring the interesting PSR. Consequently, the combination of a misleading probe set result and differential gene expression may create a false prediction of alternative splicing, or alternative start or stop (Figure 28). However, the majority of these false positives are filtered out automatically in the analysis of the exon microarray data, and most of the remaining cases can be revealed manually by inspecting the exon-wise plots.

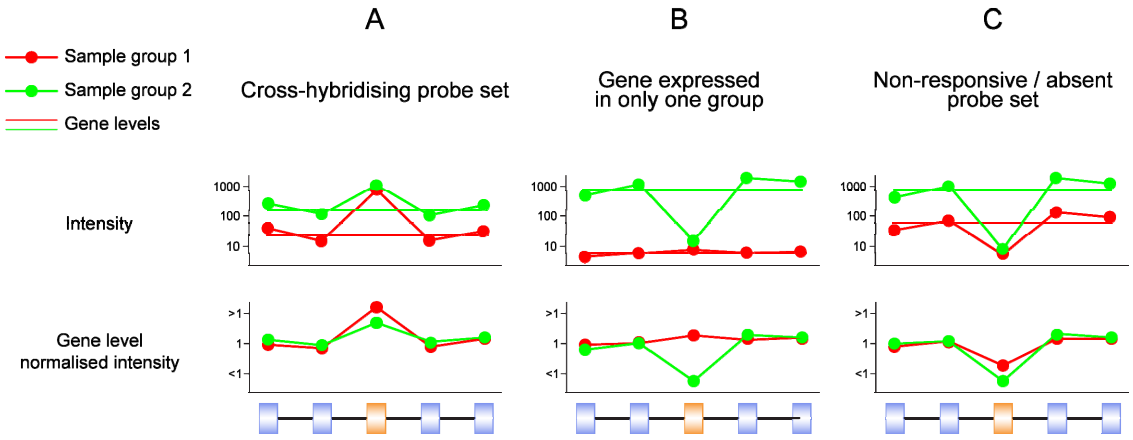


Figure 28. Scenarios that may lead to divergent transcript profiles and consequently false predictions of alternative splicing events. In each case, it is the combination of a misleading probe set result and differential gene expression that creates a false prediction of alternative splicing. **(A)** A cross-hybridising probe set giving rise to a high signal regardless of the actual expression of the interrogated PSR is depicted. Consequently, this PSR will appear to have a greater inclusion in the sample group that has a lower expression of the cognate gene. All interesting probe sets were visualised and inspected in the X:Map Genome Browser, which gave warnings when searching for probe sets known to be cross-hybridising. Such candidate events were not included in the validation pipeline. **(B)** and **(C)** A dysfunctional probe set may produce a false evidence for differential splicing in cases where the gene is expressed in only one of the sample groups and in cases with differential gene expression, respectively. Consequently, candidate genes that are not expressed in both sample groups, in addition to PSRs not expressed in at least one sample group, should be filtered out.

Candidate genes consistent with known alternative splicing, or alternative start or stop events, are more likely to be true positives. Nevertheless, in this thesis, candidate events representing possible unknown transcriptome diversifying events were not excluded, as this would have made the discovery of novel transcripts impossible. However, none of these candidates were validated, and hence these hypothetical events lowered the validation frequency.

Generally, the predicted splicing events that remain after different types of filtering are more likely to be true positives. However, it should be noted that filtering the data in an attempt to reduce the amount of false positives generally also involves loss of true positives.

In general, it was found that the smaller the p-value, and the greater the magnitude of difference between the two groups, the greater the probability of a true positive candidate.

6. Conclusions

During this study, we have established a pipeline for experimental validation of transcript variations detected by high resolution exon-level transcriptome analyses. We have shown that in total about 20 % of such variation can be validated although this may vary between test series. A transcript variant of *ZNF195* is shown to be promising as a cancer specific marker in a stem cell context. Furthermore, a transcript variant of the *SLC39A14* is shown to be a promising novel biomarker for CRC with 98 % sensitivity as compared to normal mucosa samples.

7. Future studies

To elucidate the potential clinical applicability of the presented candidate cancer-specific transcripts, we need to examine these in larger clinical series, including more normal colorectal cancer tissue samples. Furthermore, especially in the case of *SLC39A14* alternative splicing variants, we would like to test a series of adenomas to explore whether a potential diagnostic test could be sensitive to precursor lesions. The novel biomarkers could be tested for in various biological specimens, *e. g.* faeces and blood. Preferably, a test at the protein level would be sought. In most of our successfully validated transcripts, as is the case for *ZNF195*, a section of the transcript is lacking in the cancer samples as compared to the corresponding normal samples. Thus, commercially available antibodies are unlikely to target the cancer specific exon-exon junction epitope. A possibility would be to custom-order antibodies targeting this region, and apply these immunohistochemically to tissue microarrays of colorectal carcinomas and TGCTs already constructed in our department.

Biomarker applications do not necessarily require functional insights on the specific variants. However, identified cancer-specific transcripts may as well shed light on biological processes connected to cancer development, and not at least, they may serve as potential drug targets. The functional consequences of cancer-specific transcript variants will be examined by performing RNA interference based on short interfering RNAs (siRNA) or short hairpin RNA in cell lines with identified expression of the particular variant. These small RNAs can silence the expression of specific transcript variants by either post-transcriptional silencing or transcriptional silencing [74]. Recently, it has been shown that siRNAs targeting intronic or exonic sequences close to an alternative exon can regulate the splicing of that exon by local heterochromatin formation and consequently a slow-down in RNA polymerase II elongation, which may result in exon inclusion (recall from section 1.2.1) [75].

In the present thesis, transcript variants from four genes have been introduced as potential cancer biomarkers. As relatively few of the CRC study candidate genes were examined in the lab, and the proportion of validated genes was quite high, it may be advantageous to add more of these candidates into the validation pipeline. Furthermore, in our group, exon microarrays are now performed on more samples of both CRC tissue samples and EC and ES cell lines, and subsequent analysis of these data will certainly propose more candidates worthy further examination.

8. Reference list

1. Nowell PC: **The clonal evolution of tumor cell populations.** *Science* 1976, **194:** 23-28.

2. Heim S, Teixeira MR, Dietrich CU, Pandis N: **Cytogenetic polyclonality in tumors of the breast.** *Cancer Genet Cytogenet* 1997, **95:** 16-19.

3. Esteller M: **Epigenetics in cancer.** *N Engl J Med* 2008, **358:** 1148-1159.

4. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100:** 57-70.

5. Gruber SB, Ellis NA, Scott KK, Almog R, Kolachana P, Bonner JD, Kirchhoff T, Tomsho LP, Nafa K, Pierce H, Low M, Satagopan J, Rennert H, Huang H, Greenson JK, Groden J, Rapaport B, Shia J, Johnson S, Gregersen PK *et al*.: **BLM heterozygosity and the risk of colorectal cancer.** *Science* 2002, **297:** 2013.

6. Nowell PC: **Tumor progression: a brief historical perspective.** *Semin Cancer Biol* 2002, **12:** 261-266.

7. Weinberg RA: **In Retrospect: The Chromosome trail.** *Nature* 2008, **453:** 725.

8. Nowell PC, Hungerford DA: **Chromosome Studies on Normal and Leukemic Human Leukocytes.** *J Natl Cancer Inst* 1960, **25:** 85-109.

9. Rowley JD: **A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukemia identified by Quinacrine Fluorescence and Giemsa Staining.** *Nature* 1973, **243:** 290-293.

10. Hehlmann R, Hochhaus A, Baccarani M: **Chronic myeloid leukaemia.** *Lancet* 2007, **370:** 342-350.

11. Ren R: **Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia.** *Nat Rev Cancer* 2005, **5:** 172-183.

12. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nat Med* 2004, **10:** 789-799.

13. Tsujimoto Y, Gorham J, Cossman J, Jaffe E, Croce CM: **The t(14;18) chromosome translocations involved in B-cell neoplasms result from mistakes in VDJ joining.** *Science* 1985, **229:** 1390-1393.

14. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K *et al*.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:** 860-921.

15. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M *et al*.: **The sequence of the human genome.** *Science* 2001, **291:** 1304-1351.

16. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:** 931-945.

17. Lin E, Li L, Guan Y, Soriano R, Rivers CS, Mohan S, Pandita A, Tang J, Modrusan Z: **Exon array profiling detects EML4-ALK fusion in breast, colorectal, and non-small cell lung cancers.** *Mol Cancer Res* 2009, **7:** 1466-1476.

18. Wahl MC, Will CL, Luhrmann R: **The spliceosome: design principles of a dynamic RNP machine.** *Cell* 2009, **136:** 701-718.

19. Zhou Z, Licklider LJ, Gygi SP, Reed R: **Comprehensive proteomic analysis of the human spliceosome.** *Nature* 2002, **419:** 182-185.

20. Faustino NA, Cooper TA: **Pre-mRNA splicing and human disease.** *Genes Dev* 2003, **17:** 419-437.

21. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456:** 470-476.

22. Blencowe BJ: **Alternative splicing: new insights from global analyses.** *Cell* 2006, **126:** 37-47.

23. Kornblihtt AR: **Chromatin, transcript elongation and alternative splicing.** *Nat Struct Mol Biol* 2006, **13:** 5-7.

24. Cartegni L, Chew SL, Krainer AR: **Listening to silence and understanding nonsense: exonic mutations that affect splicing.** *Nat Rev Genet* 2002, **3:** 285-298.

25. McGlincy NJ, Smith CW: **Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense?** *Trends Biochem Sci* 2008, **33:** 385-393.

26. Liu HX, Cartegni L, Zhang MQ, Krainer AR: **A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes.** *Nat Genet* 2001, **27:** 55-58.

27. Pajares MJ, Ezponda T, Catena R, Calvo A, Pio R, Montuenga LM: **Alternative splicing: an emerging topic in molecular and clinical oncology.** *Lancet Oncol* 2007, **8:** 349-357.

28. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA: **Mammalian RNA polymerase II core promoters: insights from genome-wide studies.** *Nat Rev Genet* 2007, **8:** 424-436.

29. Davuluri RV, Suzuki Y, Sugano S, Plass C, Huang TH: **The functional consequences of alternative promoter use in mammalian genomes.** *Trends Genet* 2008, **24:** 167-177.

30. Li TW, Ting JH, Yokoyama NN, Bernstein A, van de WM, Waterman ML: **Wnt activation and alternative promoter repression of LEF1 in colon cancer.** *Mol Cell Biol* 2006, **26:** 5284-5299.

31. Agarwal VR, Bulun SE, Leitch M, Rohrich R, Simpson ER: **Use of alternative promoters to express the aromatase cytochrome P450 (CYP19) gene in breast adipose tissues of cancer-free and breast cancer patients.** *J Clin Endocrinol Metab* 1996, **81:** 3843-3849.

32. Mayr C, Bartel DP: **Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.** *Cell* 2009, **138:** 673-684.

33. Bartel DP: **MicroRNAs: target recognition and regulatory functions.** *Cell* 2009, **136:** 215-233.

34. Nagaraj SH, Gasser RB, Ranganathan S: **A hitchhiker's guide to expressed sequence tag (EST) analysis.** *Brief Bioinform* 2007, **8:** 6-21.

35. Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, Schweitzer A, Awad T, Sugnet C, Dee S, Davies C, Williams A, Turpaz Y: **Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.** *BMC Genomics* 2006, **7:** 325.

36. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** *Science* 2003, **302:** 2141-2144.

37. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet* 2008, **40:** 1413-1415.

38. Looijenga LH, Oosterhuis JW: **Pathogenesis of testicular germ cell tumours.** *Rev Reprod* 1999, **4:** 90-100.

39. Skakkebaek NE: **Possible carcinoma-in-situ of the testis.** *Lancet* 1972, **2:** 516-517.

40. Skakkebaek NE, Berthelsen JG, Giwercman A, Muller J: **Carcinoma-in-situ of the testis: possible origin from gonocytes and precursor of all types of germ cell tumours except spermatocytoma.** *Int J Androl* 1987, **10:** 19-28.

41. van de Geijn GJ, Hersmus R, Looijenga LH: **Recent developments in testicular germ cell tumor research.** *Birth Defects Res C Embryo Today* 2009, **87:** 96-113.

42. Clarke MF, Dick JE, Dirks PB, Eaves CJ, Jamieson CH, Jones DL, Visvader J, Weissman IL, Wahl GM: **Cancer stem cells--perspectives on current status and future directions: AACR Workshop on cancer stem cells.** *Cancer Res* 2006, **66:** 9339-9344.

43. Hussain SA, Ma YT, Palmer DH, Hutton P, Cullen MH: **Biology of testicular germ cell tumors.** *Expert Rev Anticancer Ther* 2008, **8:** 1659-1673.

44. Atkin NB, Baker MC: **Specific chromosome change, i(12p), in testicular tumours?** *Lancet* 1982, **2:** 1349.

45. Skotheim RI, Lothe RA: **The testicular germ cell tumour genome.** *APMIS* 2003, **111:** 136-150.

46. Baker DE, Harrison NJ, Maltby E, Smith K, Moore HD, Shaw PJ, Heath PR, Holden H, Andrews PW: **Adaptation to culture of human embryonic stem cells and oncogenesis in vivo.** *Nat Biotechnol* 2007, **25:** 207-215.

47. Andrews PW, Matin MM, Bahrami AR, Damjanov I, Gokhale P, Draper JS: **Embryonic stem (ES) cells and embryonal carcinoma (EC) cells: opposite sides of the same coin.** *Biochem Soc Trans* 2005, **33:** 1526-1530.

48. Parkin DM, Bray F, Ferlay J, Pisani P: **Global cancer statistics, 2002.** *CA Cancer J Clin* 2005, **55:** 74-108.

49. Bray F: *Cancer in Norway 2007 - Cancer incidence, mortality, survival and prevalence in Norway*. http://www.kreftregisteret.no; 2008.

50. Martinez-Bouzas C, Beristain E, Ojembarrena E, Errasti J, Mujika K, Viguera N, Tejada MI: **A study on MSH2 and MLH1 mutations in hereditary nonpolyposis colorectal cancer families from the Basque Country, describing four new germline mutations.** *Fam Cancer* 2009, **8:** 533-539.

51. Genuardi M, Viel A, Bonora D, Capozzi E, Bellacosa A, Leonardi F, Valle R, Ventura A, Pedroni M, Boiocchi M, Neri G: **Characterization of MLH1 and MSH2 alternative splicing**

and its relevance to molecular testing of colorectal cancer susceptibility. *Hum Genet* 1998, **102:** 15-20.

52. Goncalves V, Theisen P, Antunes O, Medeira A, Ramos JS, Jordan P, Isidro G: **A missense mutation in the APC tumor suppressor gene disrupts an ASF/SF2 splicing enhancer motif and causes pathogenic skipping of exon 14.** *Mutat Res* 2009, **662:** 33-36.

53. Kinzler KW, Vogelstein B: **Lessons from hereditary colorectal cancer.** *Cell* 1996, **87:** 159-170.

54. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP: **CpG island methylator phenotype in colorectal cancer.** *Proc Natl Acad Sci U S A* 1999, **96:** 8681-8686.

55. Søreide K, Janssen EA, Soiland H, Korner H, Baak JP: **Microsatellite instability in colorectal cancer.** *Br J Surg* 2006, **93:** 395-406.

56. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, Kang GH, Widschwendter M, Weener D, Buchanan D, Koh H, Simms L, Barker M, Leggett B, Levine J, Kim M, French AJ, Thibodeau SN, Jass J, Haile R *et al*.: **CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer.** *Nat Genet* 2006, **38:** 787-793.

57. Dukes CE: **The classification of cancer of the rectum.** *The Journal of Pathology and Bacteriology* 1932, **35:** 323-332.

58. Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP: **FIRMA: a method for detection of alternative splicing from exon array data.** *Bioinformatics* 2008, **24:** 1707-1714.

59. Obuse C, Iwasaki O, Kiyomitsu T, Goshima G, Toyoda Y, Yanagida M: **A conserved Mis12 centromere complex is linked to heterochromatic HP1 and outer kinetochore protein Zwint-1.** *Nat Cell Biol* 2004, **6:** 1135-1141.

60. Wang Y, Devereux W, Stewart TM, Casero RA, Jr.: **Cloning and characterization of human polyamine-modulated factor-1, a transcriptional cofactor that regulates the transcription of the spermidine/spermine N(1)-acetyltransferase gene.** *J Biol Chem* 1999, **274:** 22095-22101.

61. Palakurthy RK, Wajapeyee N, Santra MK, Gazin C, Lin L, Gobeil S, Green MR: **Epigenetic silencing of the RASSF1A tumor suppressor gene through HOXB3-mediated induction of DNMT3B expression.** *Mol Cell* 2009, **36:** 219-230.

62. Aleman A, Cebrian V, Alvarez M, Lopez V, Orenes E, Lopez-Serra L, Algaba F, Bellmunt J, Lopez-Beltran A, Gonzalez-Peramato P, Cordon-Cardo C, Garcia J, del Muro JG, Esteller M, Sanchez-Carbayo M: **Identification of PMF1 methylation in association with bladder cancer progression.** *Clin Cancer Res* 2008, **14:** 8236-8243.

63. Seiler N, Raul F: **Polyamines and apoptosis.** *J Cell Mol Med* 2005, **9:** 623-642.

64. Hussey DJ, Parker NJ, Hussey ND, Little PF, Dobrovic A: **Characterization of a KRAB family zinc finger gene, ZNF195, mapping to chromosome band 11p15.5.** *Genomics* 1997, **45:** 451-455.

65. Hartmann TB, Mattern E, Wiedemann N, van DR, Willemze R, Niikura T, Hildenbrand R, Schadendorf D, Eichmuller SB: **Identification of selectively expressed genes and antigens in CTCL.** *Exp Dermatol* 2008, **17:** 324-334.

66. Skotheim RI, Lind GE, Monni O, Nesland JM, Abeler VM, Fossa SD, Duale N, Brunborg G, Kallioniemi O, Andrews PW, Lothe RA: **Differentiation of human embryonal carcinomas**

in vitro and in vivo reveals expression profiles relevant to normal development. *Cancer Res* 2005, **65:** 5588-5598.

67. Shitashige M, Satow R, Honda K, Ono M, Hirohashi S, Yamada T: **Increased susceptibility of Sf1(+/-) mice to azoxymethane-induced colon tumorigenesis.** *Cancer Sci* 2007, **98:** 1862-1867.

68. Shin S, Rossow KL, Grande JP, Janknecht R: **Involvement of RNA helicases p68 and p72 in colon cancer.** *Cancer Res* 2007, **67:** 7572-7578.

69. Taylor KM, Morgan HE, Johnson A, Nicholson RI: **Structure-function analysis of a novel member of the LIV-1 subfamily of zinc transporters, ZIP14.** *FEBS Lett* 2005, **579:** 427-432.

70. Vallee BL, Auld DS: **Zinc coordination, function, and structure of zinc enzymes and other proteins.** *Biochemistry* 1990, **29:** 5647-5659.

71. Addink NWH: **A possible correlation between the zinc content of liver and blood and the cancer problem.** *Nature* 1950, **166:** 693.

72. Liu Y, Zhu X, Zhu J, Liao S, Tang Q, Liu K, Guan X, Zhang J, Feng Z: **Identification of differential expression of genes in hepatocellular carcinoma by suppression subtractive hybridization combined cDNA microarray.** *Oncol Rep* 2007, **18:** 943-951.

73. Thorsen K, Sorensen KD, Brems-Eskildsen AS, Modin C, Gaustadnes M, Hein AM, Kruhoffer M, Laurberg S, Borre M, Wang K, Brunak S, Krainer AR, Torring N, Dyrskjot L, Andersen CL, Orntoft TF: **Alternative splicing in colon, bladder, and prostate cancer identified by exon-array analysis.** *Mol Cell Proteomics* 2008, **7:** 1214-1224.

74. Castanotto D, Rossi JJ: **The promises and pitfalls of RNA-interference-based therapeutics.** *Nature* 2009, **457:** 426-433.

75. Allo M, Buggiano V, Fededa JP, Petrillo E, Schor I, de la MM, Agirre E, Plass M, Eyras E, Elela SA, Klinck R, Chabot B, Kornblihtt AR: **Control of alternative splicing through siRNA-mediated transcriptional gene silencing.** *Nat Struct Mol Biol* 2009, **16:** 717-724.

Appendix I – Primer and probe information

Gene	Probe set(s)	Name	Type	Sequence	Modification 5'	Modification 3'	Study
ABCC3	3726746	ABCC3_ex24_F_6FAM	f	GTGGATGCCAACCAGAGAAG	6-FAM		CRC
ABCC3	3726746	ABCC3_ex29_R	r	CTCTGAGTAGCTGCCGAAGG			CRC
ABCC3	3726746	ABCC3_ex24_F_TM	F	GAGAAGCTGCTACCCCTACATCA			CRC
ABCC3	3726746	ABCC3_ex25_R_TM	R	ACGAACTCCACTCCGATGCT			CRC
ABCC3	3726746	ABCC3_ex24/25_P	P	CTCCAACCGGTGGCT	6-FAM	MGB, NFQ	CRC
ARSB	2863982	ARSB_ex7_F	f	GCTCCAGCAAAGGATGACTC			TGCT
ARSB	2863982	ARSB_ex8_R	r	AGCCACAGTCATGGTGCAG			TGCT
ASPH	3137569, 3137566	ASPH_ex16_F_6FAM	f	CGCAAATACCCTCAGAGTCC	6-FAM		CRC
ASPH	3137569, 3137566	ASPH_ex19_R	r	GCCTTCAGGATGAAGCCATA			CRC
ASPH	3137569, 3137566	ASPH_ex20_R_6FAM	r	TCTTTGTTCCCAACCCTCTG	6-FAM		CRC
ASPH	3137569, 3137566	ASPH_ex17/18_F	f	TTTCTAGGTCATATGAGAGGTTCC			CRC
ASPH	3137569, 3137566	ASPH_ex10_F	f	AACATACCAAGTCTATGAGGAACAA			CRC
ASPH	3137569, 3137566	ASPH_ex9_F	f	AAGATGAAAGATTGCACCATGA			CRC
ASPH	3137569, 3137566	Hs00944865_m1	a	Unknown	6-FAM	MGB, NFQ	CRC
C2orf33	2530554	C2orf33_ex3_F_6FAM	f	TGCTAGTGTGATAATGCAAGTTCC	6-FAM		TGCT
C2orf33	2530554	C2orf33_ex5_R	r	GACCGCTCTCTTTTTAGTCTGC			TGCT
CTNNB1	2618981	CTNNB1_ex15_F_6FAM	f	CTCCAGGTGACAGCAATCAG	6-FAM		TGCT
CTNNB1	2618981	CTNNB1_ex16_R	r	CAAGCAAGGCTAGGGTTTGA			TGCT
DDX17	3960648	DDX17_ex11_F_6FAM	f	GGCACCCATCCTTATTGCTA	6-FAM		CRC
DDX17	3960648	DDX17_ex12_R	r	GCCTCTTCCAGCACTTTGAT			CRC
DDX17	3960648	DDX17_int11-12_F_6FAM	f	CAGGCTTCTCCCAACAGAGT	6-FAM		CRC
DNMT3B	3882060, 3882062	DNMT3B_ex19_F	fs	TACTTCTGGGGCAACCTACC			TGCT
DNMT3B	3882060, 3882062	DNMT3B_ex19_F_6FAM	f	TACTTCTGGGGCAACCTACC	6-FAM		TGCT
DNMT3B	3882060, 3882062	DNMT3B_ex22_R	rs	ATGCCTTCAGGAATCACACC			TGCT
DNMT3B	3882060, 3882062	DNMT3B_ex19_F_TM	F	GATGCCATCAAAGTTTCTGCTG			TGCT
DNMT3B	3882060, 3882062	DNMT3B_ex22_R_TM	R	TGGACACGTCTGTGTAGTGCAC			TGCT
DNMT3B	3882060, 3882062	DNMT3B_ex19/22_P	P	ATGAACAGGATCTTTG	VIC	MGB, NFQ	TGCT
DNMT3B	3882060, 3882062	DNMT3B_ex20/21_P	P	TAGCCAAGTTAAAGAAAG	6-FAM	MGB, NFQ	TGCT
EXOSC8	3485880	EXOSC8_ex4_F_6FAM	f	TCTGCTTTAGTGAAGTTGGGAAA	6-FAM		TGCT
EXOSC8	3485880	EXOSC8_ex8_R	r	CATCGTAGTCGAGGCAAATG			TGCT
GCNT3	3596161	GCNT3_ex1_F_6FAM	f	TGCTGAAGGGAAACAGATGA	6-FAM		CRC

GCNT3	3596161	GCNT3_ex2_R	r	AGGAACAGTTTCAGGCTCCTC			CRC
HLTF	2700225	HLTF_ex7_F_6FAM	f	TTTCTGAGAAGGACCGACCA	6-FAM		TGCT
HLTF	2700225	HLTF_ex9_R	r	TCCATCTGCCTTTTCACTGG			TGCT
HSD17B4	2825746	HSD17B4_ex4_F	f	TTGTGAAGACAGCCCTGGAT			TGCT
HSD17B4	2825746	HSD17B4_ex7_R	r	CGTGTCACTTGGAATGAACC			TGCT
ING4	3442194	ING4_ex1_F	f	GGATCGGAAGTTGCTTTGTT			TGCT
ING4	3442194	ING4_ex4_R	r	CTTTTTGCCTTTGCTGGAAG			TGCT
ITGA5	3456736	ITGA5_ex27_F	f	AGTTGCATTTCCGAGTCTGG			TGCT
ITGA5	3456736	ITGA5_ex29_R	r	CATAGCTGCCTTCTGCCTTG			TGCT
JUN	2415098	JUN_ex1_F_6FAM	f	CCACGCAAGAGAAGAAGGAC	6-FAM		TGCT
JUN	2415098	JUN_ex1_R	r	AGGGAGCGCAGGGTTAAT			TGCT
NANOG	3403437	NANOG_ex1_F	f	TCCTTGCAAATGTCTTCTGCT			TGCT
NANOG	3403437	NANOG_ex3_R	r	TGCGTCACACCATTGCTATT			TGCT
NR4A1	3415268	NR4A1_ex7_F	f	CACAGCTTGCTTGTCGATGT			TGCT
NR4A1	3415268	NR4A1_ex8_R	r	ATGGGGTGGCATATGAGAGT			TGCT
PBRM1	2676224	PBRM1_ex25_F_6FAM	f	GTGGGGACAGAATGGAGAAA	6-FAM		TGCT
PBRM1	2676224	PBRM1_ex29_R	r	ATATGGAGGTGGTGCCTGCT			TGCT
PMF1	2361418	BGLAP_ex3_F_6FAM	f	AGCTGTCTTGAATGCCTTGG	6-FAM		TGCT
PMF1	2361418	BGLAP_ex4_R	r	GTTCTCGGCCTCCTGTTTCT			TGCT
PMF1	2361418	BGLAP_ex3_F_TM	F	AAGAAGGCAAAGTCCGCAAA			TGCT
PMF1	2361418	BGLAP_ex4_R_TM	R	GGCCTCCTGTTTCTGCACAT			TGCT
PMF1	2361418	BGLAP_ex3/4_P	P	AGCCTGCAACGGGA	6-FAM	MGB, NFQ	TGCT
PTPRU	2327826	PTPRU_ex3_F(2)	f	ATCCTGCTTCTCAGCTACCC			TGCT
PTPRU	2327826	PTPRU_ex7_R	r	ACGTTGTAGCCCAGTGGTTC			TGCT
SF1	3377080, 3377078, 3377077	SF1_ex2_F_6FAM	f	CCCAAGTAAGAAGCGGAAGA	6-FAM		Both
SF1	3377080, 3377078, 3377077	SF1_ex1_F_6FAM	f	GAGGCTTGCGAAGGAGAAG	6-FAM		Both
SF1	3377080, 3377078, 3377077	SF1_ex3_R	r	GTGCGCAGTTTACGAGTCAG			Both
SF1	3377080, 3377078, 3377077	SF1_ex2'_R(2)	r	AAAATGCAGATCTTGCTCAGAA			Both
SLC39A14	3089375, 3089381, 3089382	SLC39A14_ex3_F_TM	F	GGCCAAGCGCTGTTGAAG			CRC
SLC39A14	3089375, 3089381, 3089382	SLC39A14_ex5_R_TM	R	TCTTCCAGAGGGTTGAAACCAA			CRC
SLC39A14	3089375, 3089381, 3089382	SLC39A14_ex4'_P	P	CTCACTGATTAACCTGGCC	6-FAM	MGB, NFQ	CRC
SLC39A14	3089375, 3089381, 3089382	SLC39A14_ex4_P	P	ACCGTCATCTCCCTCTG	VIC	MGB, NFQ	CRC
SLC6A6	2611941	SLC6A6_ex12_F_6FAM	f	TTGTGGGTTGCATTCTTTGA	6-FAM		CRC
SLC6A6	2611941	SLC6A6_ex14_R	r	TTGACGAGCGAGAAGATGAA			CRC
SMARCAD1	2736280	SMARCAD1_ex5_F_6FAM	f	CAAGCACTATGGATGGAGCA	6-FAM		TGCT
SMARCAD1	2736280	SMARCAD1_ex9_R	r	CCATTTGGAACCTCACTTTG			TGCT
SNRPN	3584495	SNRPN_ex3_F_6FAM	f	TGGAGTAGCGAGGAATCTGA	6-FAM		TGCT
SNRPN	3584495	SNRPN_ex5_R	r	CATCTTGCAGGATACATCTCATT			TGCT

SPAG9	3762591	SPAG9_ex2_F_TM	F	GCGAAAAACTATGCTGACCAGA			TGCT
SPAG9	3762591	SPAG9_ex3'_F_TM	F	GTGGCAGCAAGCTGAATGAC			TGCT
SPAG9	3762591	SPAG9_ex5_R_TM	R	TCTCCAGCAGGTAATGGGAAA			TGCT
SPAG9	3762591	SPAG9_ex3/4_P	P	CACACTGAGATGATCCAT	VIC	MGB, NFQ	TGCT
SPAG9	3762591	SPAG9_ex3'_P	P	CTGCTGTTTGTGTTTGG	6-FAM	MGB, NFQ	TGCT
TARBP2	3416118	TARBP2_ex1_F	f	CTCGTCGGCTGTGTATTGG			TGCT
TARBP2	3416118	TARBP2_ex2_R	r	CAGGCGTCTTCCCTATTCTG			TGCT
TDGF1	2620943	TDGF1_ex1_F	f	CGATGCTAACGCCTCTTTTC			TGCT
TDGF1	2620943	TDGF1_ex3_R	r	GAGATGGACGAGCAAATTCC			TGCT
UBA6	2771755	UBA6_ex20_F_6FAM	f	TGAAGTTATTGTACCGCATTTGA	6-FAM		TGCT
UBA6	2771755	UBA6_ex23_R	r	CAGCCTTCTAAACTGTGTCCA			TGCT
UCK1	3227651	UCK1_ex4_F_6FAM	f	GTTTGAGGGCATCTTGGTGT	6-FAM		TGCT
UCK1	3227651	UCK1_ex7_R	r	TCAGAATGTCCTGGATGTGC			TGCT
ZNF195	3359764	ZNF195_ex4_F	f	TGGGATTTCACCATGCTACTC			TGCT
ZNF195	3359764	ZNF195_ex4_F_6FAM	f	TGGGATTTCACCATGCTACTC	6-FAM		TGCT
ZNF195	3359764	ZNF195_ex6_R	r	CTGGCAGAAGGTCTTGGGTA			TGCT

Abbreviations: f, forward RT-PCR primer; r, reverse RT-PCR primer; F, forward real-time RT-PCR primer; R, reverse real-time RT-PCR primer; P, TaqMan MGB probe; a, TaqMan gene expression assay (contains both primers and probe); fs, forward RT-PCR and sequencing primer; rs, reverse RT-PCR and sequencing primer.